2

ULTRA PRECISION MACHINING
ONR FINAL REPORT

Contract # N00014-89-J-1608

1982-1990

Principal Investigators:
Prof. Daniel B. DeBra, Dept. of Aeronautics and Astronautics
Prof. Lambertus Hesselink, Dept. of Aeronautics and Astronautics
Prof. Thomas Binford, Dept. of Computer Science

Stanford University
Stanford, CA 94305-4035

May 20, 1990

60

Contributors to ONR Research

The principle contributors who have worked on and been supported by the research program covered by this report are listed below. In addition, our development of techniques for ultra precision machining has stimulated projects in a number of courses which are given after the list of people directly involved in the project.

SENSING

Faculty	*Bert Hesselink*
Students	*Steve Collicott*

ACTUATION

Faculty	*Daniel DeBra*
	Dave Beach
Staff	*Ed Ditzen*
	Russ Hacker
	Dick VanPatten
Students	*Cliff Oostman*
	Chingling Chou
	Chien-Jen Chen
	Hy Tran
	Graham Ross
	J.C. Tsai

COMPUTING

Faculty	*Tom Binford*
Students	*Arni Geirsson*
	John Bourg

There are three courses that have benefited directly. These include: a course in Precision Engineering (ME 119), our Introductory Course in Manufacturing Technology (ME 103) and a course in Fluid Power Control (ME 229). The closest match has been between the Precision Engineering course and our research on the machine tool. Since 1983, the projects in the Precision Engineering course have included: 1) A preliminary design of a laminar flow motor by Mike

King, 2) Straight edge and the kinematic mounting for straight edge reversal by Jen-Kou Chou, 3) The air suspension system including the development of a Laminar Flow Restrictor for Damping for the Hydraulic Test Bed by Mark McCullough, John Howard and Richard Warner, 4) A precision spindle metrology instrument for evaluating all six degrees of freedom of the motion of a rotating shaft. The initial measuring head work was by Pat McCune and Leslie Howard. Work in parallel on the electronics associated with the capacitive pick-offs was done by Chingling Chou. Subsequent work on the circuit was carried out by Sang-Il Lee and Soo Hong Lee. An elastic suspension for fine adjustments of the metrology head was done by Haim Kedar. 5) A short stroke carriage mounted on an elastically suspended Watt mechanism was carried out by Jeff Chung and Karen Bennett. 6) An elastic hinge lever device for attenuating motion by a factor of 300 for the purpose of calibrating LVDT's with a micrometer was developed by Richard Warner and a subsequent design by Kurt Ohms. 7) An actuator for making small motions that would be used to evaluate our magnetic read out for the spindle angle on the machine tool and therefore for calibrating Hall effect detectors used in that application was developed by Mustafa Sayed. 8) In the early development of the machine tool base work we were experimenting with artificial granite. This included the evaluation of whetl.er the artificial granite would be a satisfactory medium with which to make air bearings without inserts. George Tsai developed some prototypes using molding techniques and experimented with different combinations of artificial granite and molding release agents to provide us some guidelines which eventually steered us away from considering that technology. We also had some student development of artificial granite material which developed thermal and long term creep strain evaluations and gave some indications of porosity under both air and fluids. As it turned out the surface is impervious. There is a skin forms against the form but if this gets removed then the artificial granite structure is porous and can allow air to pass through and/or absorb fluids which would compromise the stability of these materials in the presence of coolants and cutting fluids.

In ME 113 a number of less ambitious projects were completed e.g. a student made sliding plexiglass doors to shield the operator from oil spray when cutting.

The fluid power control course (ME229) was enriched with lecture material, homework and test problems that came directly from the design of the quiet hydraulic components.

Please note that all sections are self contained in regards to figures and references.

Contents

Chapter 1

INTRODUCTION

In 1982, we received funding from the US Office of Naval Research to pursue three areas in support of Precision Manufacturing. We had identified metrology, actuation and the computational support of diamond turning machines as target areas for our contribution. The work was funded under ONR contract # N00014-84-J-1608 for six years and continued last year under ONR contract # N00014-89-J-1608.

There are a number of fields that require or can use to advantage very high precision in machining. For example, further development of high energy lasers and x-ray astronomy depend critically on the manufacture of light weight reflecting metal optical components.[Robinson, C.A., 1972; Wills-Moren et al, 1982]

To fabricate these optical components with machine tools ~~they~~ mirrors will be made of metal with mirror quality surface finish. By "mirror quality surface finish" we mean the dimensional tolerances on the order of 0.02 ~~μm~~ ~~(1 μin.)~~ and surface roughness of 0.07 ~~μm~~ ~~(3 μin.)~~. These accuracy targets fall in the category of "ultra precision machining". They cannot be achieved by a simple extension of conventional machining processes and techniques [Tonaguchi, N., 1983; Brown, N.J. et al, 1983]. They require single crystal diamond tools, special attention to vibration isolation, special isolation of machine metrology, and on line correction of imperfection in the motion of the machine carriages on their ways.

Metrology is a key aspect of both monitoring and correcting the imperfections in a machine tool. The first problem that we have addressed in machine tool metrology (section 2) was an improved technique for measuring the non-straightness of ways using the same laser pathway employed for carriage displacement measurement. Whether this laser pathway is evacuated, uses a helium atmosphere or works directly in air but shielded from convection, it is critical in the machine design that this space be made available in the appropriate operational locations. Once there, using it for non-straightness measurement simplifies the design of the machine. The first section of the body of this repor (section 2) on sensing describes the technology that has been developed.

Actuation for existing diamond turning machines has been through the use of electrical motors and hydraulics. We have felt that hydraulics has an opportunity to provide better temperature control, however conventional hydraulics have developed to provide high performance and efficiency in a small package. The resulting devices are unsuited for use on an ultra precision lathe due to the flow and pressure fluctuations, so we have developed a family of actuation devices based entirely on laminar flow which avoid the problems of conventional hydraulic devices. Accepting lower efficiency and larger size in return for smooth and quiet operation, we call the approach "quiet hydraulics". Section 3 of this report describes the components developed and the temperature control that has been developed to provide a suitable control fluid.

A computing environment for a modern sophisticated machine tool may play many roles not thought of in conventional numerical control. It can provide an historical record of the decision on design, modification and maintenance throughout the history of the machine. It can provide an on-line design support for modifying the numerical description of a part in situ if necessary. It can provide monitoring for machine faults or degradation of performance and provide helpful support in diagnosing these and providing corrective action. Calibration data and techniques can be available on-line when needed and normal operating procedures and recommended operating conditions can be made available. It would be enormously ambitious to incorporate all these possibilities in the supporting computational environment for a machine tool in one step. We have chosen the monitoring of the health of a machine tool as the first step. This work is described in section 4.

REFERENCES

Robinson, C.A., 1972, "Defense Dept. Backs Space-Based Missile Defense". *Aviation Week and Space Technology*, Sept. 12, pp. 14-16.

Wills-Moren, W.J., Modjarrad, H., and Read, R.F.J., 1982, "Some Aspects of the Design and Development of a Large High Precision CNC Diamond Turning Machine".*Annals of the CIRP*, Vol.31/1, pp.409-414.

Taniguchi. M. 1983, "Current Status in, and Future Trends of, Ultraprecision Machining and Ultrafine Materials Processing". *Annals of the CIRP*, Vol.32/2, pp.573-582.

Brown, N.J., and Donaldson, R.R., 1983, "Fabrication of Machined Optics for Precision Applications". *Proceedings of SPIE*, Vol.32/2, pp.48-62.

Chapter 2

SENSING

2.1 INTRODUCTION

The goal of this research is to develop an optical technique for real-time high-precision measurements of tool-bit positions. Optical technology and systems developed for this purpose are equally capable of measuring a solid body strain or displacement field or an instantaneous two-dimensional fluid velocity field in real-time. Thus the development of one of these instruments is quite equivalent to another in the optical processing sense. Some differences arise with the need for operating in different environments, but these minor adaptations involve the construction of a specific system, and not the technology involved. Thus the research for high-precision shear interferometry has concentrated on the application of real-time speckle velocimetry. The technology and knowledge derived from this application benefits shear interferometry, speckle metrolgy, and speckle velocitmetry.

Of course, both "instantaneous" and "real-time" are relative terms, and are used somewhat loosely here because of the exploratory nature of the optics research. For machine tool applications, the time scales involved would be those of tool motion and cutting speeds, and also tool velocities when moving between cutting operations. An instantaneous fluid velocity field would be a velocity field which was sampled in a time period shorter than the fastest characteristic time scale of the motion. Real-time measurement of this instantaneous two-dimensional velocity field would present numerical values for the velocity field within the time between samples of the velocity field. Obviously, an instrument used to measure a real-time instantaneous velocity field in a liquid convective cell would not be capable doing so for a hypersonic flow, so some specific scale (spatial and temporal) of flow had to be chosen. Our goal has been to make these real-time instantaneous velocity field measurements in a gaseous or liquid flow which is of high enough Reynolds number to be of interest in turbulence studies.

The starting point for our research was the relatively new optical technique known as laser speckle velocimetry. As existing at the beginning of our research, laser speckle velocimetry could be used to record an instantaneous two-dimensional velocity field on photographic film. After chemical development, this photographic record could be analyzed with one of two different optical processors to measure either the velocities or the iso-velocity contours throughout the recorded region of the fluid flow.

An instrument to perform the above operations in real-time could be used to measure tool-bit position and velocities on a lathe or other machine tool. Such an instrument would also be an useful tool for the study of the basic physics of gaseous, liquid, and solid materials.

2.2 METHOD

Our approach to designing a real-time laser speckle velocimetry system concentrated on improving the temporal operation of three portions of the overall system: optical recording of speckle patterns, optical processing of the recorded information, and digital measurement of the output of the optical processor. In implementing speckle metrology and velocimetry in real-time, the optical recording and optical processing portions operate at the same time, but are best described separately. Progress in these three areas is described below.

2.2.1 Photorefractive Crystals

We have demonstrated the real-time (35 microseconds) speckle pattern recording capabilities of photorefractive crystals in an optical instrument which can measure an instantaneous velocity field of a solid body in motion. Application of this technology to measuring an instantaneous two-dimensional fluid velocity field has proven substantially more difficult and is discussed below, along with an alternate real-time recording medium which is under consideration.

The simplest real-time laser speckle velocimetry system is shown in Figure 2.1 . A spinning disk may appear to be a trivial object to measure, but it is an ideal object for studying the capabilities of this new optical instrument; it has a known two-dimensional velocity field containing all directions and an easily controlled magnitude. A double- pulsed Neodymium:YAG (Nd:YAG) laser is used for illumination. The two pulses from this laser are less than 10 nanoseconds (10-8 seconds) in length and the time between pulses can be varied between 35 and 250 microseconds. The laser light is scattered by the rough surface of the disk (this experiment could also use reflected, rather than transmitted light), and is imaged into the photorefractive crystal. Two pulses produce two exposures, with the second exposure shifted relative to the first because of the motion of the object.

The photorefractive crystal is Bismuth Silicon Oxide, Bi12 SiO 20, which is commonly called BSO. Prior to the start of our research, this (and similar) photorefractive materials had been studied by many groups and numerous holographic recording experiments had been conducted. Our use of photorefractive crystals was unique (and may still be) because the crystal is used to record speckle patterns produced by an imaging system, not holograms, and the time between exposures is shorter than any previously reported double exposure work. We have studied the response of the crystal to the double exposure speckle patterns2 and found that with a large enough energy density (exposure), the crystals can record the double exposure speckle patterns.

A photorefractive crystal absorbs a portion of the light incident on the crystal during exposure, just as photographic film absorb the light incident on it. A semi- conductor model is generally used to describe the crystal's response to the illumination. Each photon which is absorbed may excite an electron from a trap sight into the conduction band. Once in the conduction band, the electron moves under the influence of both an applied electric field and diffusion. In the speckle recording experiments, electron transport due to the applied field is the dominant mechanism. The electrons fall back into trap sights (where the electrons are stationary) at a rate which is characteristic of the material. For the photorefractive crystals used (BSO and BGO) this average recombination time is about 5 microseconds, but was found to be as long as 300 microseconds and as short as microseconds in some samples of the material (this large variation between samples is typical of the current state of the art in photorefractive crystal manufacturing.) Once the electrons are trapped, the charge distribution within the crystal is inhomogeneous. This stationary space-charge field, through the linear electro-optic effect, alters the propagation of light through portions of the crystal. In the simplest description, the incident speckle pattern has produced a variation in the index of refraction of the crystal. This index field mimics the incident speckle field. Two of these speckle recordings may be made in succession, resulting in a recording of double-exposure speckle patterns. For time between exposures at least ten times the average recombination time, the energy densities in the two exposures should be equal for best performance (cleanest fringes in the output plane of the optical processor.) When the time between exposures is comparable to the recombination time, the second exposure must have a larger energy density than the first exposure.

Once the simplest case had been demonstrated, we attacked the problem of measuring an instantaneous two-dimensional fluid velocity field in real time. The main difference between this and the previous, solid velocimetry, is that the fluid is seeded with small (1 micron) particles which scatter the incident laser pulses (Figure 2.2). It is the scattered light which is imaged into the crystal, and the displacements of the particles between exposures which is measured. The scattering process is a very inefficient process; most of the incident light is scattered in the forward direction, and only a small portion is scattered into all other directions. Then only a small region of the angular scatter is collected by the imaging system. Nevertheless, this general process has proven very useful in the past for flow visualization, laser Doppler velocimetry, particle-tracking experiments, and similar fluid mechanics experiments. The low photo-sensitivity of the photorefractive BSO crystal has prevented us from measuring an instantaneous fluid velocity field in real-time. We have determined that to produce a useful recording in these crystals, an energy density on the order of 1 milliJoule per square centimeter is required. This is analogous to a photographic film with an ASA number of 0.001, which would be an extremely slow film. Simply increasing the laser energy is neither a practical nor an economical solution; the light pulses produced by our Q- switched Nd:YAG laser must be used with great care to avoid destroying lenses, mirrors, and crystals. The response of the crystal is enhanced by applying a large DC electric field. We have found that for our crystal and experiment, a safe maximum field strength is 7.5 kV/cm. This low photosensitivity is not a problem for solids applications, such as tool-bit measurement, since the illuminating laser light is reflected or transmitted directly into the imaging system. A large fraction of the incident light is collected by the imaging lens and exposes the crystal.

Since the light scattered nearly forward by the particles is much more intense than the light scattered normal to the illumination, (Figure 2.3), we have imaged the near-forward scattered light into the crystal. Since the object plane is no longer perpendicular to the optical axis, the image is also tilted obliquely to the optical axis. Such an imaging system is termed "Scheimpflug" imaging after the man who first wrote down the relations between object and image tilt (in the mid-nineteenth century). We have shown that the displacements between the two speckle patterns may still be measured by optical processing. Collecting this near-forward scattered light increased the energy density incident on the crystal by an order of magnitude, but was still insufficient to produce a detectable recording. This type of oblique (or Scheimpflug) imaging is applicable to traditional speckle velocimetry using photographic film recording, and allows a ten times larger region of the flow field to be recorded without an increase in the laser power.

The scattering efficiency of small particles was then studied in order to optimize this portion of the experiment. Particle sizes from 0.01 to 10 microns diameter (silica, smoke, and aluminum oxide) and seeding densities from 10^3 to 10^{11} particles per cubic millimeter were studied. The only instance when the energy density incident on the crystal is close to large enough is when the seeding density is increased to an absurd level. At such a high level, the scattered light collected by the imaging system has been scattered by several particles as it passes out of the flow, and therefore is not useful for speckle velocimetry. It is doubtful if a fluid with 10^{11} particles per cubic millimeter actually behaves as a Newtonian fluid. Oblique imaging and optimizing the seeding particles are two methods to increase the magnitude of the energy density, and hence the "strength" of the recording and the magnitude of the signal produced by the optical processing of the recording.

We also examined one method to detect smaller signals from the optical processor, in hopes of reducing the minimum exposure required to produce a measurable recording in the photorefractive crystal. Since the photorefractive BSO is both optically active and birefringent (with applied electric field), the polarization state of the light which comprises the signal is different from the polarization state of the background light. This property allows us to filter the output of the optical processor (Figure 2.4) in order to detect a smaller signal than without the filtering.

Several different photorefractive crystals were studied, including Bismuth Germanium Oxide, Barium Titanate, and Strontium Barium Niobate. The first is so similar to BSO that the two can be considered as identical for this experiment. Barium Titanate and Strontium Barium Niobate have been used by others in countless continuous illumination experiments, but proved unsuitable for our high-speed pulsed-illumination needs. Barium Titanate and Strontium Barium Niobate are excellent materials for continuous illumination experiments, but fail miserably in short-pulsed experiments. Both of these materials have a very short average electron lifetime in the conduction band, and hence a very short ($< 1micron$) average electron transport difference. This short transport length is not a problem in continuous illumination experiments, for the trapped electron can be re-excited into the conduction band. For short-pulsed illumination, such as our experiments, the pulse length is much shorter than the average electron lifetime. Hence each electron is excited only once, and is transported only once. It turns out that an average transport length of less than one micron is insufficient to record speckle patterns. In speckle patterns, the smallest length scale is one the order of five or ten microns, and often times larger.

One might specify desirable properties for an ideal crystal for this application, or even simply specify what improvements to BSO would be needed to permit real-time laser speckle velocimetry in a fluid flow. However, the basic physics of the photorefractive process are still not fully understood. While theories describing electron transport within the crystal have been proposed and are very useful in many experiments, it is unclear how to alter the crystal growth processes, the purity of the crystal, the dopants or the material composition to create certain desired properties. Currently we are engaged in quantifying the requirements for a new material or an improved BSO so that as new photorefractive materials or variations of existing ones are developed, we will know whether or not they are suitable for the fluid velocimetry application.

This research has demonstrated for the first time that photorefractive crystals such as BSO and BGO are practical materials for real-time speckle metrology applications. The speckle patterns have been recorded with as little as 35 microseconds between the two exposures. The Q-switched pulses from the Nd:YAG laser are three orders of magnitude shorter than the time between exposures, so any blurring of the speckle patterns during exposure is negligible.

2.2.2 Optical Processing

Optical processing is used to measure the velocities recorded in the real-time recording medium. The same optical processors used to measure the velocities recorded on photographic film can be used to measure the velocities recorded in the photorefractive crystals. However, an new application of anamorphic (non-axisymmetric) optics developed early in our work provides for multiple-point optical processing in real-time.

Original studies of the strength and quality of the signal produced by optical processors operating on speckle patterns have produced a theory to predict signal strengths from volume speckle recordings as found in photorefractive crystals described above. This same theory was found to be an accurate model for thinner recordings of speckle, such as in photographic film. In either case, the amount of the interrogating light diffracted into the signal (diffraction efficiency) and the ratio of the signal energy to the energy of the primary noise source (transmitted diffraction efficiency) are predicted. The quality of the signal improves dramatically as the mean optical density increases, to the point where there is no need to be block the undiffracted spike in order to clean up the recorded fringes. The maximum diffraction efficiency for film recordings is 14% at a mean optical density of 0.49, but the transmitted diffraction efficiency is only 43% . This shows that the strongest signal is not necessarily the best signal. Maximum diffraction efficiency produces the highest intensity in the output plane. This may be important in some applications in order to minimize exposure time for recording the fringes, or to maximize a weak signal. When these two considerations are not important, the quality of the recorded fringes may be improved dramatically by increasing the exposure used to record the speckle patterns. Although maximum intensities in individual speckle patterns are generally not known exactly, and nonlinearities of the photographic recording process affect the statistics of recorded speckle patterns, the assumption that the spatial diffraction rate is the maximum peak-to-peak change in absorption length appears to be useful for mean absorptions up to half of the maximum absorption of the film. For a speckle size of about twelve times the free-space interrogating wavelength, and an emulsion thickness of twenty wavelengths, the thin absorption speckle model is found to be suitable for predicting the diffraction efficiency, defined as the ratio of signal power to incident power, and the transmitted diffraction efficiency, defined as the ratio of signal power to transmitted power. Both efficiencies are predicted accurately by considering the absorption of the signal wave within the absorbing film emulsion. The need for such analysis is demonstrated by the failure of thin-plate predictions. The results of the thin absorption speckle analysis predict that the maximum diffraction efficiency from absorption speckle is found for a mean optical density of 0.48, but that the signal-to-noise ratio increases monotonically with the mean optical density. This implies that the cleanest signal is found for higher mean optical densities, limited ultimately by the finite dynamic range of the recording material. Hence the proper exposure for maximum transmitted diffraction efficiency is that exposure which is just short of saturating the film.

An optical processor measures the displacements by interfering the Fourier transforms of the first and second exposures. Since one of the speckle patterns is shifted relative to the other, the Fourier transforms arrive at the output plane travelling slightly different mean directions. The difference between directions results in interference fringes in the output plane. This interference can occur with either two- or one-dimensional Fourier transforms. In the anamorphic processor (Figure 2.5), one-dimensional Fourier transforms are interfered to produce fringes related to one velocity component in the flow. A one-dimensional imaging system with a one-dimensional image plane coincident with the one-dimensional Fourier transform plane permits all the points in one line through the flow to be processed in parallel. The result is a set of curved fringes which are represent the spatial distribution of velocities in the flow. This is commonly called a "velocity profile", and is of interest in many fluid mechanics studies. It is important to note that no scanning or film transport is required in this multipoint process, and as described below, the digital image processing routines used to measure the fringe pattern are inherently faster than those used to measure the spacing and orientation of a fringe pattern produced by a two-dimensional Fourier transform optical processor. This anamorphic optical processor is applicable to double exposure speckle patterns recorded in either photographic film or photorefractive crystals.

This should be considered as a method which is complimentary to two-dimensional Fourier transforming three and two-dimensional spatial filtering of the recorded speckle patterns. These three optical processors operate on the same multiple-exposure speckle patterns, but produce different measurements. The anamorphic system is ideal for measuring and displaying a single-component instantaneous velocity or displacement profile. Two-dimensional Fourier transforming measures both components of the recorded displacement, but only at one point in the field. Two-dimensional spatial filtering operates on the entire field at one time and produces a contour map of velocities, but not an absolute measurement of velocities. All three processors are unable to distinguish the sign of the displacement vectors, i.e., directional ambiguity is inherent to the processors and must be removed in the recording process. One observation made following publication of the original paper describing the application of an anamorphic processor to measuring speckle pattern displacements is that the fringe modulation diminishes and ultimately vanishes as the mean velocity or displacement vector deviates from the direction of the one-dimensional Fourier transform. The cross-flow component of the displacement reduces the strength of the signal relative to the noise, resulting in less accurate velocity measurements, or no measurements at all in the extreme case. The study of this loss of modulation has led to the analysis and experiments presented here. Given a double exposure speckle pattern in the input plane, the intensity in the output plane is determined, and the fringe modulation is computed from the intensity. The dependence of fringe modulation on the magnitude of the recorded displacements, and on the design of the recording and processing optics is found in analytical form. The analysis predicts a certain variation of fringe modulation, and this is compared to experimental results. The results of this analysis show how to design an anamorphic optical processor which is more tolerant of cross flow, i.e., the processor can measure velocities over a larger range of cross-flow velocities. Analysis of the anamorphic optical processor begins at the input plane, where the double-exposure speckle pattern is placed. This may be a photographic film recording, or a photorefractive crystal in the case of real-time speckle velocimetry. The statistics of the double-exposure speckle patterns depend upon the imaging system used to record the speckle patterns. For recording on film, the recording and processing optics are distinct systems, but in real-time speckle velocimetry the two optical systems may share components as shown in Figure 2.1. This analysis shows how to design an anamorphic processor to best suit the flow field under study, and also provides quantitative predictions of the off-design performance capabilities. This theory predicts the effects of various components and dimensions of the anamorphic processor on fringe modulation. The predictions of fringe modulations are verified by experiments. Experiments also demonstrate an anamorphic optical processor with a six-fold increase in the tolerance to cross-flow displacement. Only one-dimensional digital image processing routines are required to process the fringe pattern for measuring the modulation or the spatial frequency of the fringes. This is inherently faster than performing the two-dimensional processing required to measure the spatial frequency of fringe patterns produced by a two-dimensional Fourier transform optical processor. A second advantage in efficiency is found when one considers that for an N by N recording of fringes, the two-dimensional Fourier transform processor produces two displacement measurements at the interrogated point. For an N by N recording of fringes, the anamorphic processor produces N velocity measurements. This can translate into reduced time to measure displacements throughout the recorded velocity or displacement field. We intend to publish this analysis and results within the year.

2.3 DIGITAL PROCESSING

The output of the optical processor contains the velocity information in the form of a fringe pattern, usually with substantial multiplicative noise. Numerous people have developed different hardware and software schemes for measuring the spacing and orientation of these fringes. When the anamorphic optical processor described above is used, the digital image processing routines (Fourier transform, Hartley transform, autocorrelation) involved are one-dimensional. This is inherently faster than performing the two-dimensional versions of the transforms. In anticipation of coupling real-time recording and digital processing, we studied the digital processing of fringes produced by an anamorphic optical processor and an oblique recording of double exposure speckle patterns recorded on photographic film. Such a fringe pattern and the resulting measurement of fringe spatial frequency is shown in Figure 2.6. We implemented a one-dimensional Hartley transform to compute the spatial power spectrum of the fringes, and from the peaks of the fundamental and second spatial harmonic, we determined the fundamental spatial frequency. Of course, one may Hartley transform the spatial power spectrum to compute the spatial autocorrelation, which will also give a measurement of fringe spatial frequency. This would require roughly twice the time as performing one transform, but may be less noisy. The choice of technique depends on the severity of the noise in the fringe pattern, and hence would most likely vary from experiment to experiment.

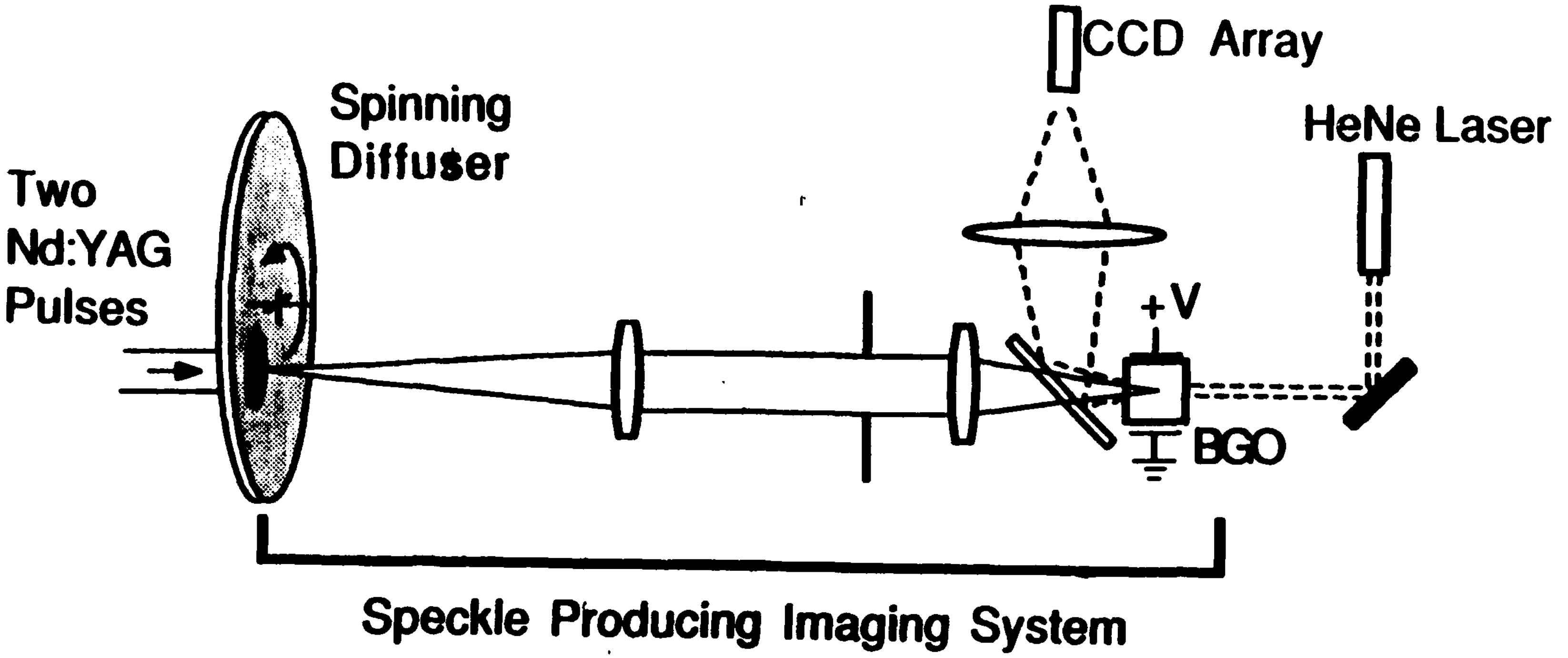


Fig. 2.1 Velocity field of spinning disk is imaged into the crystal. The Fourier transform optical processor interrogates one point. The optical processor immediately measures the displacement recorded at that point. The CCD array records the fringe pattern for digital processing.

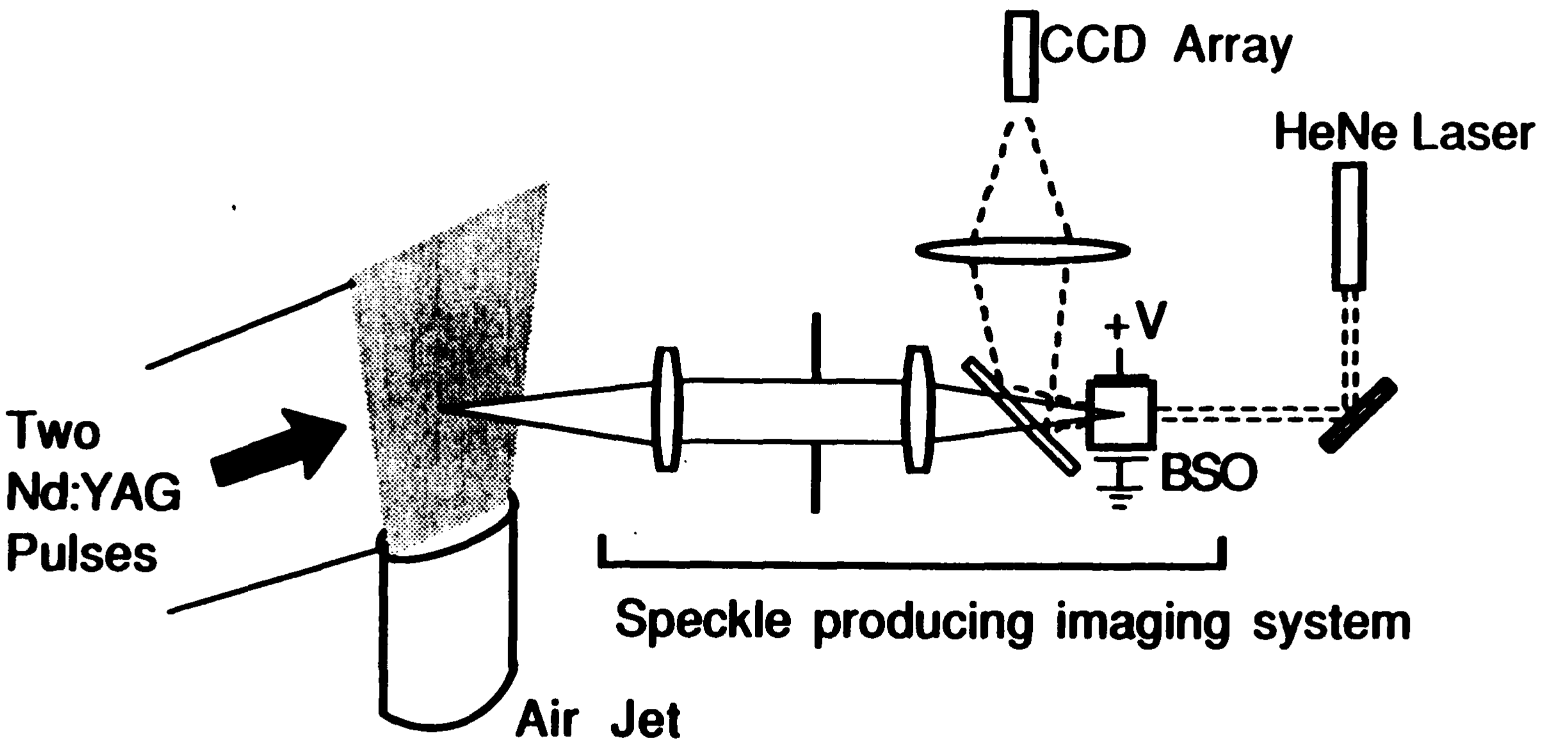


Fig. 2.2 Geometry for Real-Time Fluid Flow Speckle Velocimetry
Seed particles in the jet scatter the laser light. The scattered light is imagedinto the BSO crystal and the resulting image is a speckle pattern. The motion of the seed particles between exposures is measured by the optical processor.

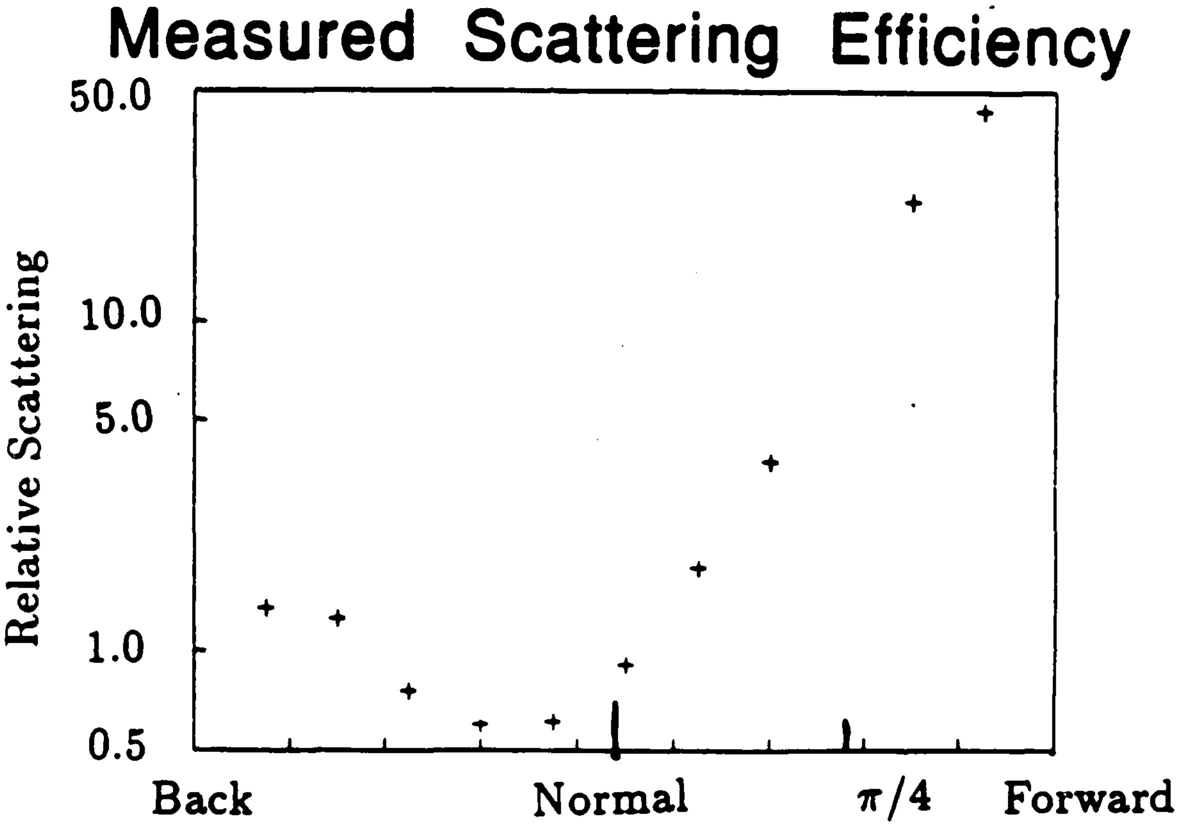


Fig. 2.3 Scattering efficiency versus scattering angle.

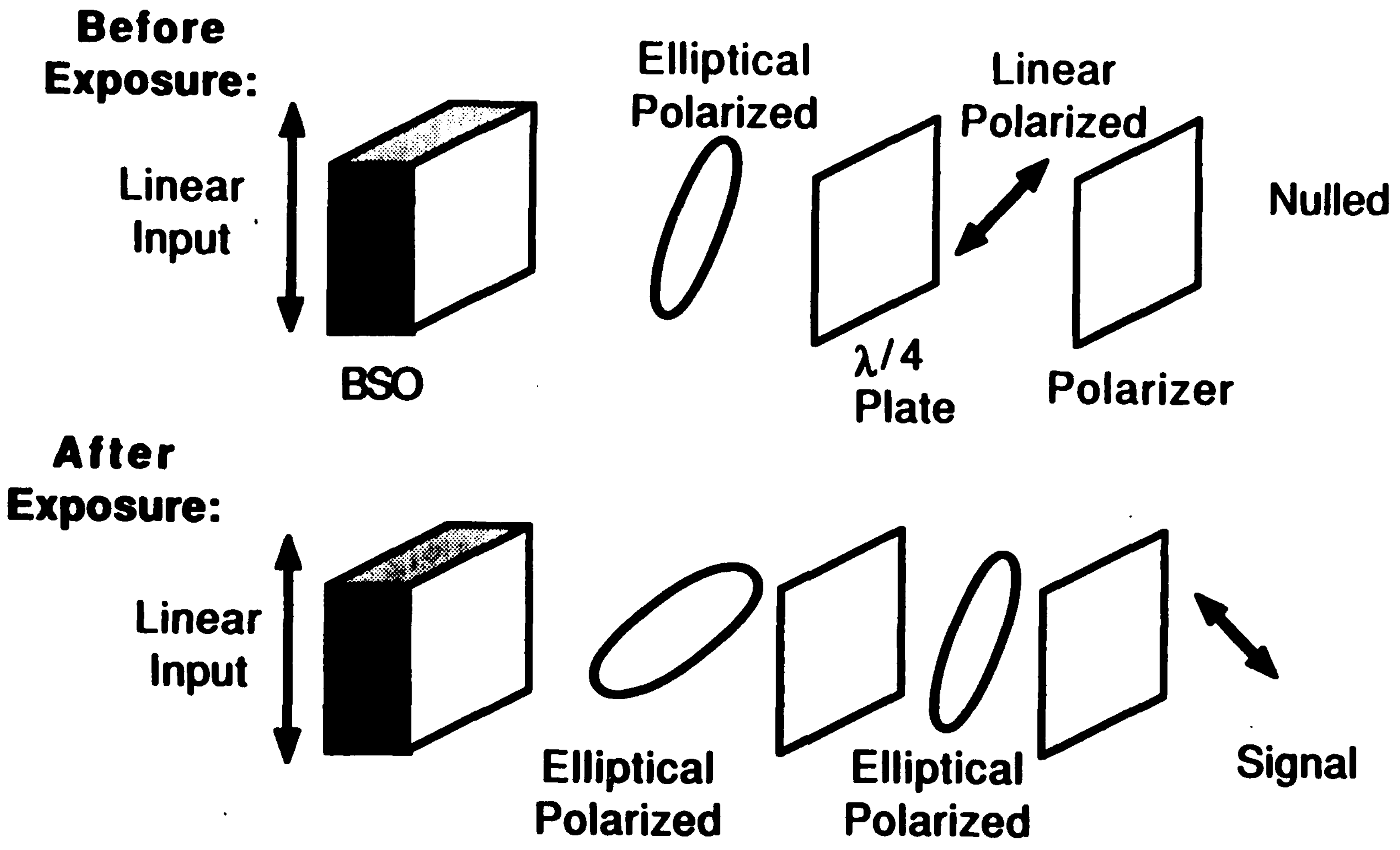


Fig. 2.4 Schematic of polarization filtering of the optical processor light transmitted by an optically active and birefringent BSO crystal.

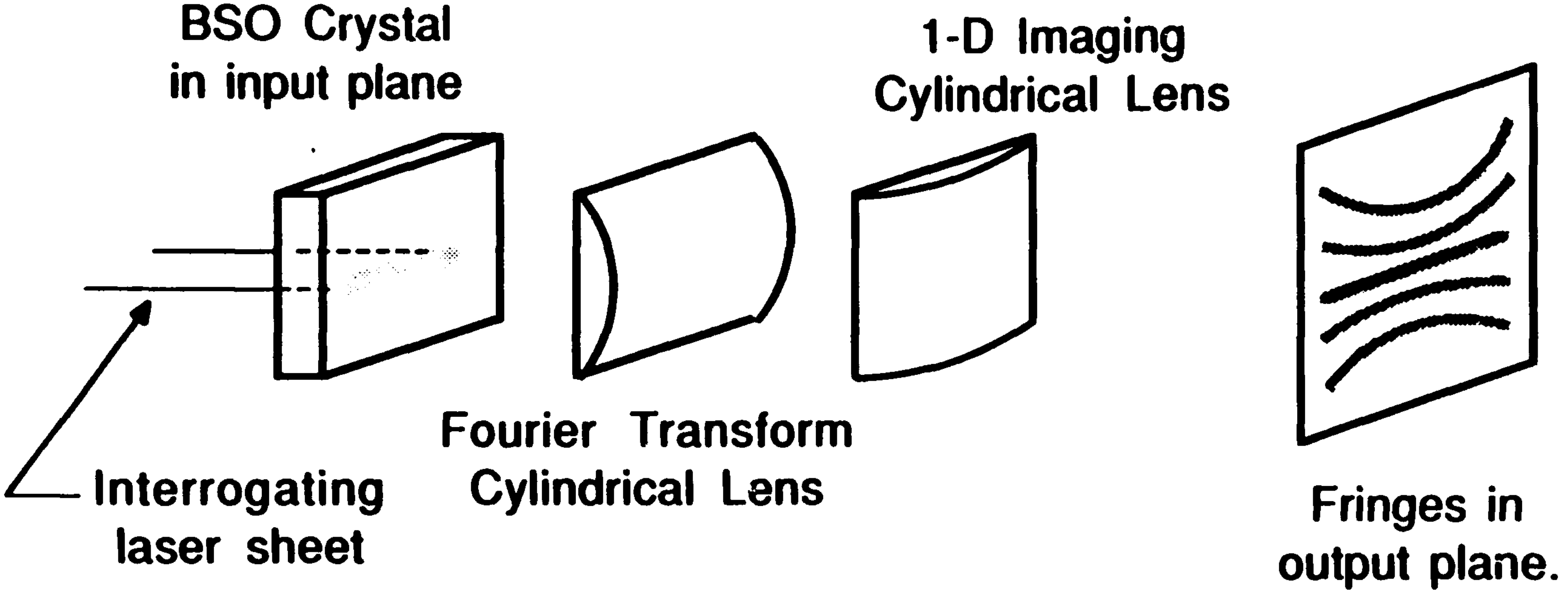
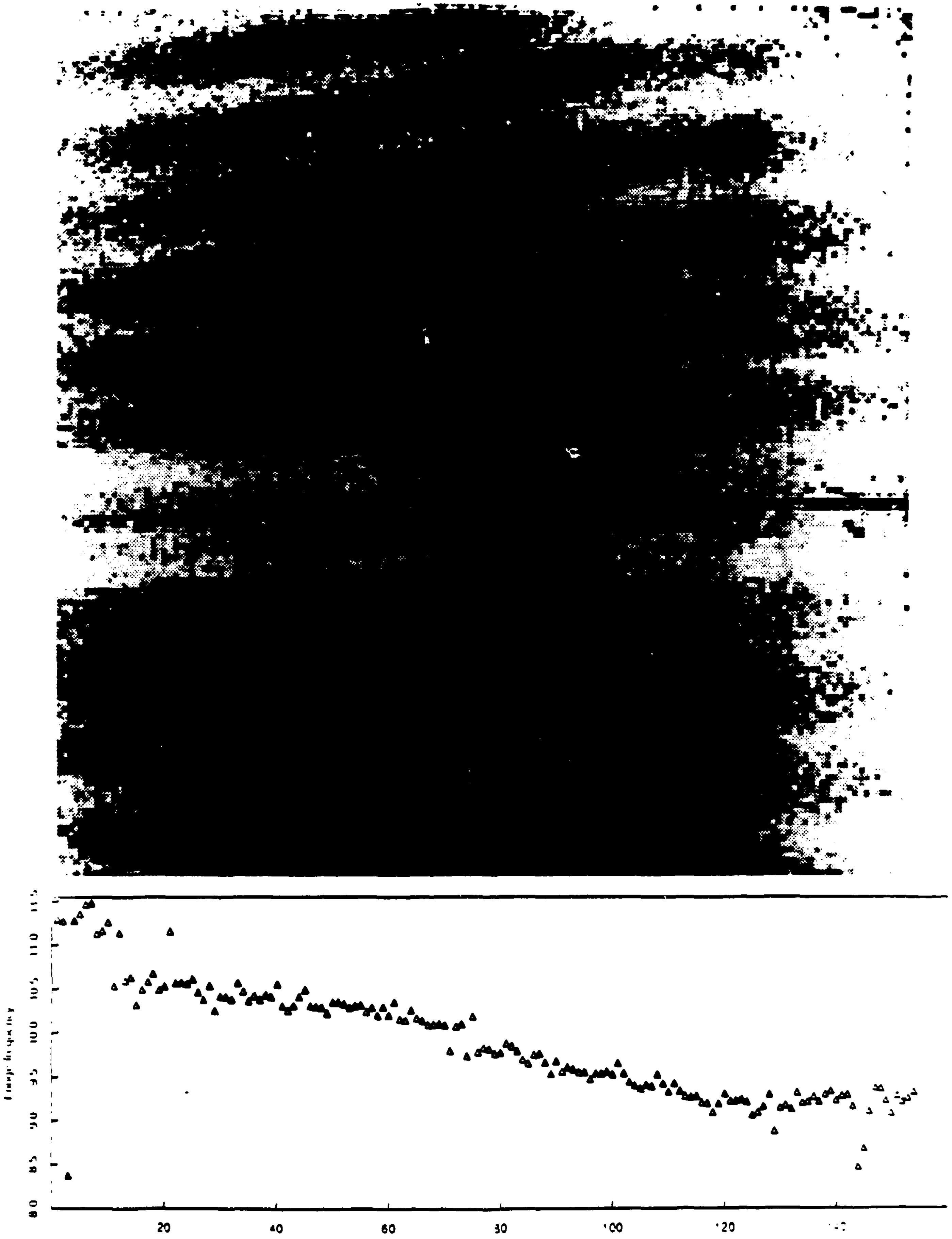


Fig. 2.5 Anamorphic Optical Processor.
The 1-D Fourier transform plane of the first lens is coincident with the 1-D image plane of the second lens. The fringe spacing can be measured using only 1-D image processing routines.



Digitized fringes and measured fringe spatial frequency.

Chapter 3

ACTUATION

3.1 QUIET HYDRAULICS

3.1.1 Introduction

The development of diamond turning machines has driven the requirements for machine tool design to new levels of accuracy. Each of the technologies that have been a candidate for actuation, measurement and control of the machine tool environment has had to be re-examined.

Temperature is one of the principal disturbances when one approaches accuracies of 0.1 to 0.01 micrometers. Electric motors have been the choice for actuation and spindle drive because of their convenience and cleanliness. However, the inefficiency of an eddy current motor which has low vibration, deposits heat in the rotating part where it is hard to remove and a permanent-magnet-rotor motors produce more vibration.

By contrast, hydraulics permits the heat of inefficiency to be carried away with the working fluid. Liquids undergo a relatively small rise in temperature as they decrease in pressure and this is directly proportional to the supply pressure. Given this comparison, we began a study of the application of hydraulics to precision machining about 1982.

Conventional hydraulics have been optimized to maximize power density and efficiency. For most applications, this has been an appropriate development. This has been accomplished by using orifice modulation in valves and seals to prevent leakage. The orifice flow is turbulent causing vibrations that affect the surface finish and the seals typically have coulomb friction which is undesirable for precision stage movement.

To avoid the seismic disturbances of turbulent flow associated with orifice modulation, we chose to develop a family of hydraulic actuation devices based on laminar flow. By increasing the viscosity of the working fluid and using close tolerances in fits, we have been able to create moving parts with no contact and with acceptably small leakage.[Viersma, 1980]

For the development and evaluation of these devices, we chose initially to build a test bed for evaluation of their performance and their freedom of seismic disturbances.

3.1.2 A Test Bed for Quiet Hydraulics

We needed a bench on which to mount our quiet hydraulic actuators. It needed to be seismically isolated and could benefit from being massive. We chose to build an artificial granite test bed, shown in Fig. 3.1.1. The three legs shown are conventional hydraulic concrete. The test bed itself is polyester granite. Polyester was chosen because it is less viscous, less toxic and less expensive, while providing adequate strength for the application we had in mind. Cliff Oostman, working with Dan DeBra, developed the geometrical design incorporating a gutter for the return of oil in what was to be an open oil system. The test bed weighed 1.25 tons.

For vibration isolation, it was mounted on three air bags. These were provided with height control valves and employed laminar flow dampers for the air flow between the air bag and a surge tank. See Fig. 3.1.2. [DeBra, 1984]. The height control valve was a modified bleeding pressure regulator with the diaphragm replaced by a mechanical actuator. This project was developed by three students, Mark McCullough, Richard Warner, Joshua Howard in a precision engineering course (ME119) at Stanford, following suggestions made by Earl Lindburg. The ways of a South Bend lathe, no longer in use, were utilized for convenient mounting of components for evaluation. In Fig. 3.1.1, the test linear actuator is shown mounted.

Temperature control following the development of Bryan [1972] was chosen. The oil for showering as well as for each of the actuators needed to be temperature controlled. Chinglain Chou [1988] developed a systematic approach to using conventional heat exchangers to obtain millidegree temperature control. He incorporated feed-forward with anticipation by placing temperature measurements upstream of the input liquids for both the oil and chill water to overcome time delays in their response. The feed-back that can be obtained by measurements at the outlet of the heat exchanger are limited in bandwidth. The principal limitation is the time delay in the heat exchanger. Chou established that millidegree temperature control can be obtained, but that there is a residual fluctuation in temperature due to the variability of the heat transfer coefficient internally in the heat exchanger.

Even if the inlet oil and water temperatures are constant, the output temperature control uncertainty is $10^{-3}4$ of the temperature differential. Thus for no input variations at this level of accuracy, the heat transfer coefficient has to be constant to 10^{-4}. Since the flow is turbulent in the heat exchanger if one wants good heat transfer, one should expect to see some flow induced fluctuations in the heat transfer coefficient and consequently in the output temperature.

Thus Chou established that variations in the inlet temperatures of oil and water could be compensated by feed forward and appropriate modeling, but also demonstrated that the final limit has to do with the high frequency portion of the unsteadiness of the heat transfer coefficient internally in the heat exchanger.

For a lathe, it is necessary to have a high quality spindle and spindle drive. The spindle bearings, using oil with external pressurization, have been demonstrated on a number of machines. However, the only hydraulic motor for diamond turning was on one of the first machines developed at Philips [Kraakman, 1970]. The vane motor used had pressure fluctuations associated with it which caused a strain that appeared as a modulation in the motion of the spindle and hence the pattern cut by the diamond tool on the part. We chose to use a different approach. Instead of a constant displacement motor, we have employed the shearing of fluid across a perfectly smooth, cylindrical stator [Chen, 1987]. Fig. 3.1.3 shows the laminar flow mounted on the test bed. The rubber hoses which help isolate the motor from external seismic disturbances introduce oil to flow over one half of the rotor through a passage 0.5 mm thick and the other set receive the oil and return it to the supply.

Fig. 3.1.4 shows the disassembled motor. The delivery of torque to the shaft is the shearing over the smooth large section of the shaft. The endplates contain conventional externally pressurized bearings and in the foreground one sees the tapered pins used to adjust the external resistance for centering the shaft.

Actuators are needed to move the tool with respect to the workpiece. Whether the ways are stacked or independent is immaterial. It was necessary to employ the hydraulic fluid for this linear relative movement in order to have a consistent use of the temperature advantage of hydraulics.

Fig. 3.1.5 shows the linear actuator test apparatus. In the foreground is the tool holder which is also a short stroke actuator. Behind it, one cylindrical way has two externally pressurized bearing blocks to provide four degrees of freedom stiff for the carriage. Behind it is a differential laminar flow valve which modulates the flow to the cylinder. It has a spherical joint mounted to the rear of the carriage, completing the fifth degree of constraint for a kinematic support of the carriage. This makes the carriage motion independent of the alignment of the two ways to first order.

This actuation system was nominally chosen to have a travel of five inches (125 mm). The evaluation of this system has not been completed.

The short stroke actuator, shown in the foreground of Fig. 3.1.5, is shown in close-up in Fig. 3.1.6. A large block of steel is cut with wire EDM to produce a the front and back a plate 8 mm thick for elastic support. A bellows is placed between the remaining portions of the block, one part of which is attached to the top and other to the bottom to provide an actuation which has a travel of about 0.2 mm. The pressure in the bellows is modulated by flapper valve shown in the left foreground. This system has a bandwidth approximately of 30 Hz and would be used for example to modulate the motion of the tool for non-axisymmetric cutting [Tran, 1990].

3.1.3 Incorporating Quiet Hydraulics in a Machine Tool

Development work on the components has been sufficiently encouraging that we have chosen to incorporate these principles in a modest demonstration machine. This machine employs a three-ton hydraulic concrete base with an oil pan mounted on top. The pan separates the precision area from the rest of the machine base, which has as its principal function seismic isolation. The components will be mounted with elastic supports [Geirsson, 1988]. The design and construction has been carried out principally by John Bourg and Arni Geirsson and has recently had the first oil flowing. We expect to incorporate some ways with modulated externally pressurized bearings which can be used for correction of straightness and yaw.

3.1.4 Conclusion

The development of quiet hydraulics for precision engineering has provided some new technology which hopefully will be useful in expanding the list of candidates available for precision machine tool design. It has also provided a wonderful challenge for students. It has been necessary for them to rethink the actuation devices through from first principles. The designs that have resulted are sufficiently different than conventional hydraulic devices that there has been very little carry-over of technology. This has been a perfect environment for graduate student development of research ideas into working hardware.

REFERENCES

Bryan, J.B. et al, "A Practical Solution to the Thermal Stability Problem in Machine Tools", *SME Technical Paper* No. MR72-138,1972.

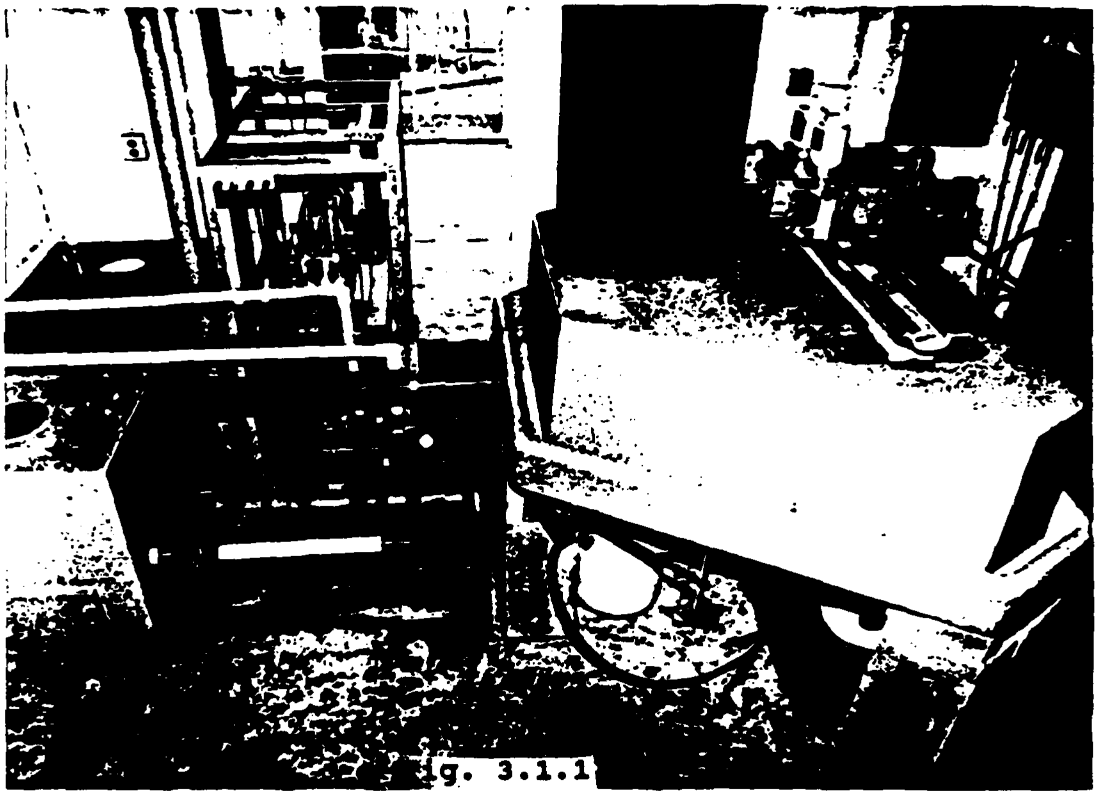
Chen, C-J, and D.B. DeBra, " A Laminar Flow Motor for Precision Machining", *CIRP*, 1987

Chou, Chinglain, "A Precise Fluid Temperature Control for Precision machines", doctoral dissertaion, Stanford Univ., 1988.

Kraakman, J.J.J. and J.G.C. deGast, "A Precision Lathe with Hydrostatic Bearing and Drive", *Philips Technical Review*, Vol. 30, No. 9, pp. 229-245, 1980.

Tran, H.D. and D.B. DeBra, "Feedforward Methods for Non Circular Turning on a Lathe", submitted for presentation at CIRP 1990 meeting.

Viersma, T.J., "Analysis, Synthesis and Design of Hydraulic Servosystems and Pipelines", *Elsevier Scientific Pub., Co.*, 1980.



ig. 3.1.1

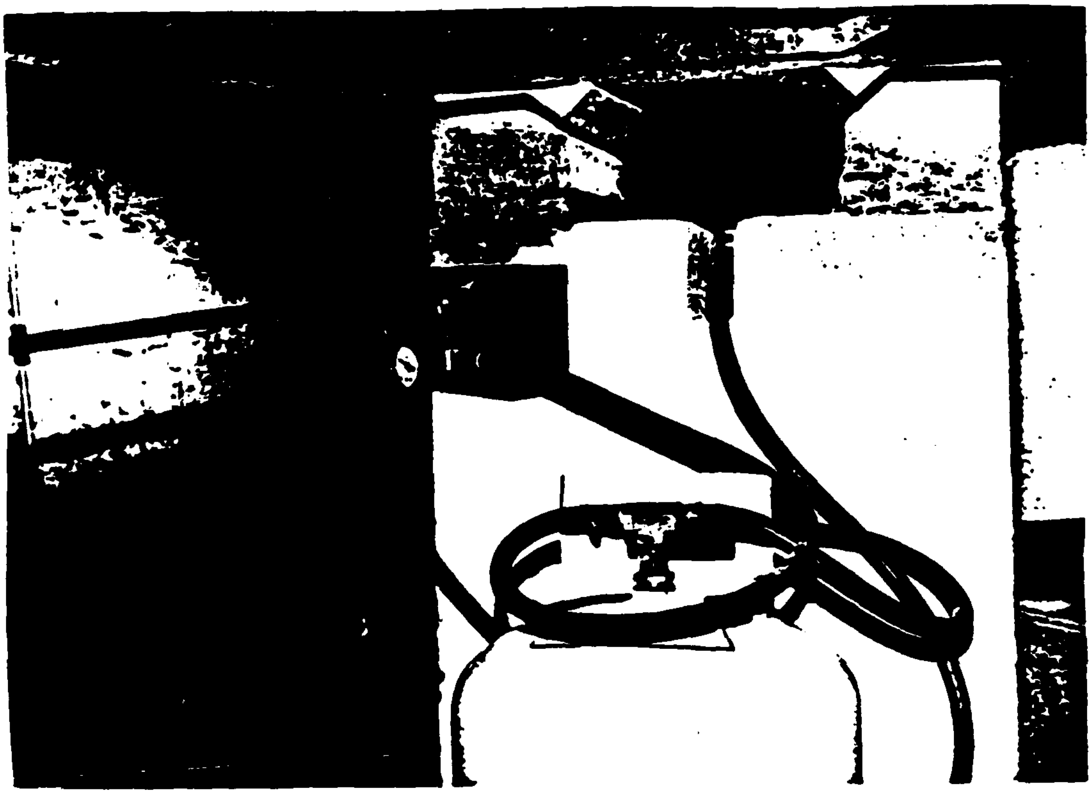


Fig. 3.1.2

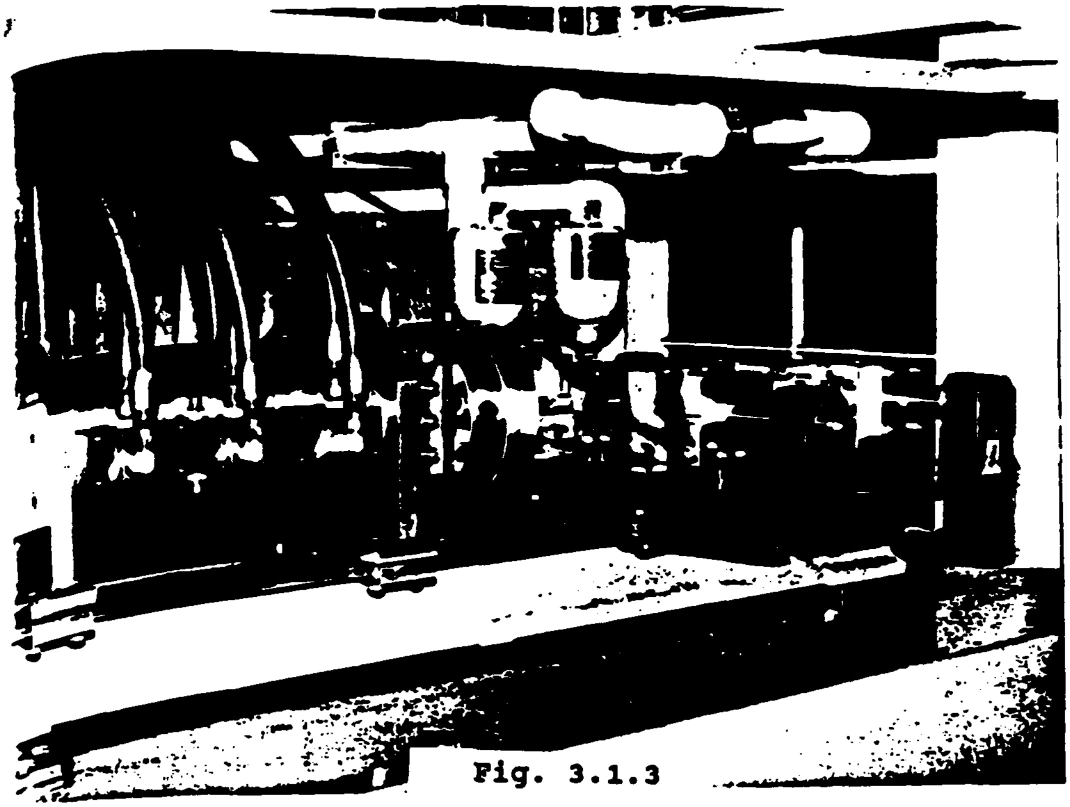


Fig. 3.1.3

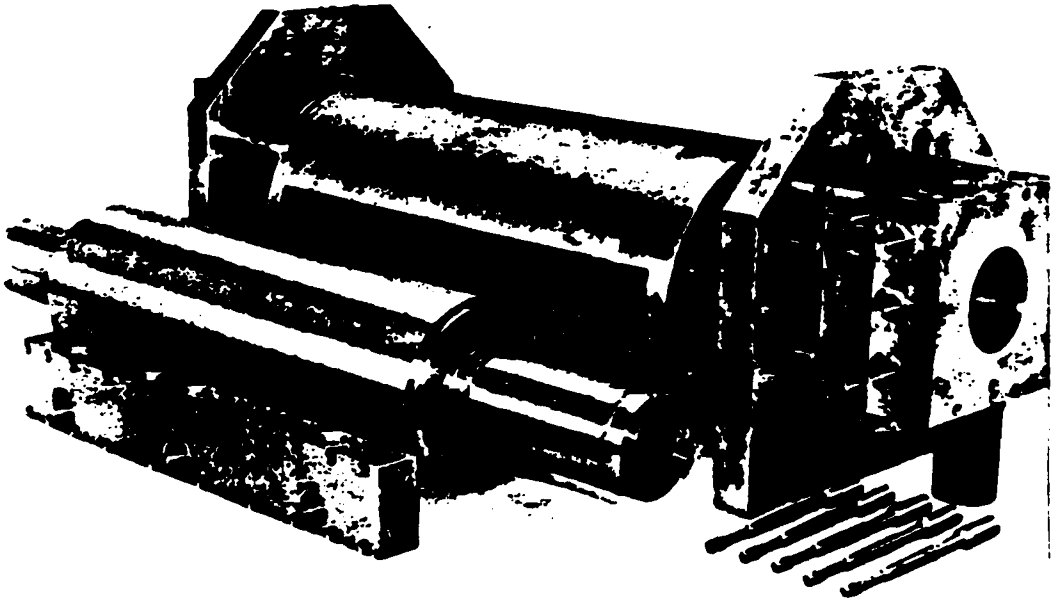


Fig. 3.1.4

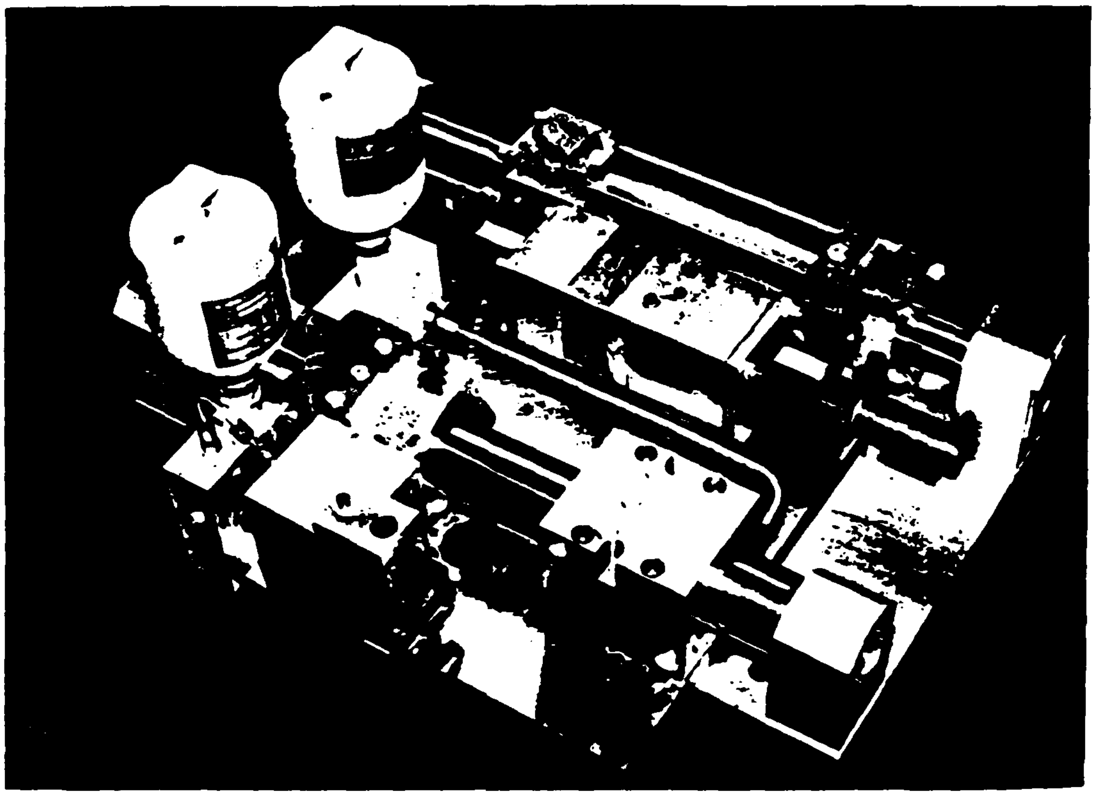


Fig. 3.1.6

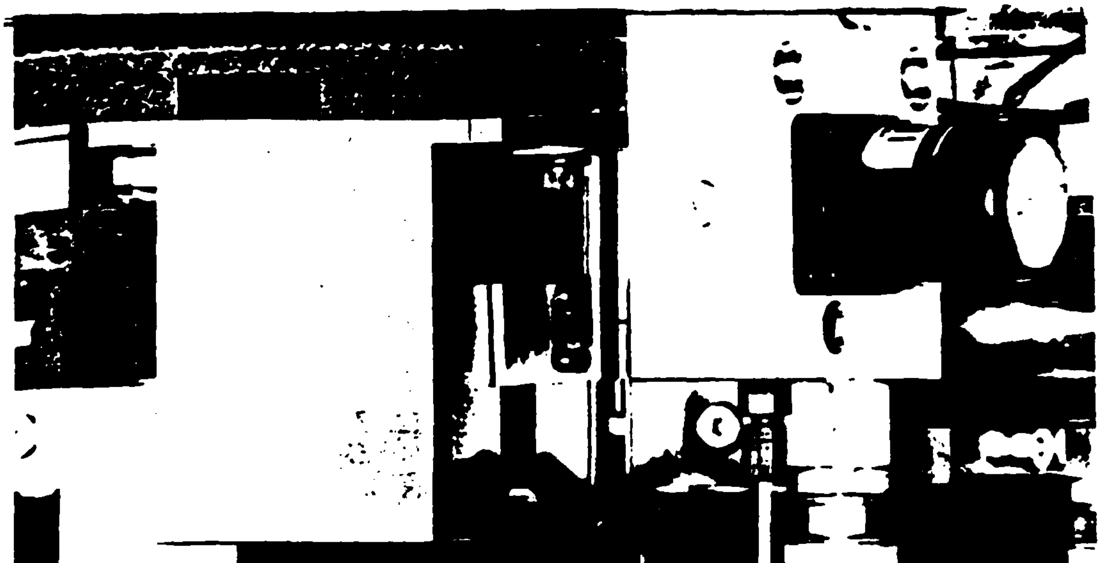


Fig. 3.1.5

3.2 LAMINAR FLOW MOTOR

3.2.1 Introduction

Diamond turning machines have been developed at a number of locations, including the Philips Research Laboratory in Holland [Kraakman, 1969; Gijsbers, 1980], Lawrence Livermore National Laboratory [Bryan, 1979; Donaldson, Patterson 1983; McCue, 1983], Oak Ridge Y-12 plan [Barkman, Woodard 1981], several industrial firms in the United States, and the Cranefield Unit for Precision Engineering in England [Will-Moren, Modjarrad, Read 1982]. There are additional machines under development in Europe and Japan and some are now commercially available.

Hydraulic actuators have the merit that heat generated due to inefficiency is carried away with the working fluid. Hydraulic fluid can be temperature controlled to the order of a millidegree [C.J. Chen, 1985] and thus improvements in temperature control of the spindle are possible, which would minimize thermal strain.

Conventional hydraulics have developed to provide high performance and efficiency in a small package. The resulting devices are unsuited for use on an ultra precision lathe due to flow and pressure fluctuations. The present device is designed to avoid these problems, accepting lower efficiency and larger size in return for smooth and quiet operation. The motor is incorporated as an integral part of the spindle.

Philips Research Laboratory began work on precision lathes around 1960. The first machine called De Gasts' [Kraakman, 1969] lathe uses hydro-static bearings and hydraulic drives to avoid the limitation of stiction. The main spindle and carriages in this lathe are supported by hydrostatic bearings such that there is no metal contact between the moving surfaces. The spindle is driven by a rotary hydraulic vane motor with 11 vanes which is fairly insensitive to overloading. A dimensional tolerance of 1 μm (40 μin.) and a maximum out-of-roundness of 0.1 μm (4 μin.) were achieved. In the early seventies, T.G. Gijsbers [Gijsbers, 1980] added computer control and an improved version of the vane motor, but an artifact of the pressure fluctuations in the vane motor remained as an eleven arm star on work pieces. This very successful early work in hydraulics encourages us to develop a motor which would operate more smoothly.

The spindle drive is shown in [Fig 3.2.1]. Flow passing over the rotor provides the torque by viscous sheer. The maximum theoretical efficiency of a laminar flow motor is derived as 33.3%. Therefore we have to minimize the friction loss in the motor and its bearings.

Leakage is unavoidable with finite clearances, but excessive leakage causes unnecessary pumping loss which also generates heat in the fluid. A power output of 250 watts at a rotating speed of 100 rad/sec (955 rpm) was chosen as the design goal for a demonstration motor.

A technology test bed for components which could be used on a diamond turning machine has been built, which is vibration isolated and supplied with oil for bearings, motor and temperature control. The base is made of artificial granite [McKeown, Morgan 1979] for thermal stability. It is isolated from ground motion by a pneumatic vibration isolation system [McKeown, Morgan 1979; DeBra, 1984; DeBra, 1981] composed of three commercially available pneumatic isolators which were modified by adding passive damping and height adjustment feedback [Warner, McCulloch, Howard 1983].

Electrical motors have been used successfully in the diamond lathes referenced earlier. Cogging and side forces vary with the type of motor and generally are higher for motors with less heat generation in the rotor and where it is difficult to provide cooling. Vibration and thermal effects have been isolated in some cases by mounting the motor off of the machine and coupling the drive through belts or a shaft with universal joints. A comparison of the drives is given in Table 1 showing the potential advantage of a laminar flow motor over other types.

As our work progressed we found references to the dual of the laminar flow motor on laminar flow pumps. These pumps were designed to avoid the pressure and flow surges of constant displacement devices. They had a long history [Modrovsky, 1957; Fazekas, 1961; Fazekas 1961; Bullough, 1974; Bardsley, 1893; Sternlicht, 1962; Fuller, 1984]. Hahn had also used viscous sheer for drive coupling in grinding [Hahn, 1965].

[Fig. 3.2.2] shows sectional views and a developed view of a laminar flow motor. The motor consists of a hollow cylindrical stator and a solid rotor which are integrated into a machine spindle. The rotor fits into the stator with a radial clearance. As the pressurized fluid flows through the clearance, the viscous shear stress acts on the periphery of the rotor. The shear stress causes the rotor to rotate in the direction of the flow. In our case, two flow paths are shown. To avoid excessive by-pass flow between the paths, very tight clearance is required. An O-ring or a doctor blade is not used in order to ensure that no dry friction is present

It is the shear stress at the surface of the rotor generated in the laminar flow which provides the driving torque. Since there is no positive displacement element such as a piston or a vane in the motor, the output torque of the motor should be ripple-free. Shear stress is generated by the velocity gradient, thus the roughness of the shearing surface of the rotor doesn't matter as long as the flow is kept laminar.

For simplicity, let us assume that all the flow supplied by a hydraulic pump is used to drive the rotor. This assumption is equivalent to saying that: 1) There is no by-pass flow in the circumferential direction between flow paths, 2) There is no leakage (side flow) in the axial direction which is perpendicular to the sliding motion, 3) The end effects can be neglected and the flow pattern is steady and fully developed, 4) The clearance h between the rotor and the stator is small compared to the other dimensions. This means $h \ll R\theta_e$ and $h \ll L$, 5) The curvature of the path is negligible $h \ll R$, where

h = the clearance between the stator and the rotor,

R = the radius of the rotor,

θ_ϵ = the effective working angle of a flow path, [Fig.3.2.2] shows : $\theta_\epsilon \simeq \frac{2\pi}{N}$

L = the length of the rotor (width of the flow path),

N = the number of paths (two are shown in [Fig. 3.2.2]).

(3.1)

This problem reduces to a laminar flow between two parallel plates which are bounded by two planes perpendicular to the flow direction.

Equation 3.2.1 defines

$$\begin{aligned} \alpha &= \frac{2\mu\omega R}{(\frac{-dp}{dx})}h^2 = \frac{1}{3} \\ (T\omega)_{max} &= \frac{1}{3}\Delta pQ = \frac{1}{3}\Delta pN\omega RhL \end{aligned} \tag{3.2}$$

At the operating condition for maximum efficiency, h can be expressed in terms of the other parameters.

$$h = \sqrt{6\mu\omega R\frac{R\theta_\epsilon}{\Delta p}} = \sqrt{\frac{12\pi\mu\omega R_2}{N\Delta p}} \tag{3.3}$$

Therefore the maximum power output is

$$(T\omega)_{max} = \frac{2\pi}{3}\Delta p^{\frac{1}{2}}N^{\frac{1}{2}}\omega^{\frac{3}{2}}R^2L\mu^{\frac{1}{2}} \tag{3.4}$$

Thus an operating point can be selected which is representative for the machine where it will have the maximum power output. One can operate at speeds below that range successfully but not as efficiently. A relatively small increase in speed is possible but one eventually reaches zero torque output as the speed increases due to the internal drag of the fluid itself.

Chen has calculated the temperature rise, friction power loss and modified the theoretical efficiency by including leakage of a laminar flow motor across the lands that are intended to keep the fluid in the channel used to provide the shearing torque. He provides an extensive discussion of all the parameters available to the designer and how they trade against each other and what practical considerations should be taken into account in designing a laminar flow motor [Chen, 1985].

3.2.2 Experimental Performance

Complete operating data are given in Chen's [1969] dissertation. [Fig. 3.2.8] shows the motor power output versus shaft speed. The design point is very close to its predicted values confirming that the simple flow models used in the theoretical development are an adequate description of this type of motor. Compromise on the oil viscosity and other factors shifted the point of optimum efficency to a lower speed as can be seen in [Fig. 3.2.9]. Our calculations of the efficiency losses relative to the optimum of a perfect motor of 331/3% predicted a maximum of 17.5%. So the drag and leakage predictions have been acceptably modeled. A more detailed comparison of theory and prediction are given in [Fig. 3.2.8] at half the nominal flow. These data were taken before the temperature contol was operating and one can see some data scatter a portion of which is due to temperature variations.

An LVDT is used to measure the random relative motion between the spindle and the tool holder. [Fig. 3.2.10] shows the spectrum of random relative motion in the radical direction. The spindle is locked such that its rotaion is constrained. The peak LVDT output value is 67.1mV corresponds to a random motion of 8.38 μin.(0.21 μm) in the radial direction. It is surprising that the dominating terms are coming from the belt, even though it is continuous with no obvious joint. To achieve a mirror surface with a diamond tool, the random motion in the radial direction must be reduced by a factor of three. Thus an accumulator for a modest amount of filtering is called for. Subsequent testing at low rpm and therefore low flow gives values of 4-5 μin. (0.1 μm) without filtering a very encouraging result.

The spectrum of the LVDT reading shows no significant peaks at high frequencies. This is an advantage of using a hydrostatic bearing. Oil has good ability to attenuate noise at high frequencies. Furthermore, dynamic stiffness at higher exciting frequencies is greater thatn the static stiffness. The higher dynamic stiffness will reduce the magnitude of error motion.

In [Fig. 3.2.11] the coherence between the LVDT readings and pressure fluctuations shows that the principal random relative vibrations of the spindle with respect to the tool holder are due to the IMO pump and its accessories, such as the check valve, reservoir, electric motor and belt. A quick analysis shows that the significant random motion of the spindle does not respond to pressure fluctuation in the bearing since it is one order of magnitude smaller than that in the laminar flow motor. Thus filtering should be quite effective in reducing the hydraulic induced random motion.

3.2.3 Spindle Performance

The symmetrical design was chosen to avoid supply pressure fluctuation from disturbing the rotor. While no special effort was made to achieve small radial runout (the measured synchronos motion is 40 μin. (1 μm), we wanted to evaluate variations frɔm revolution to revolution due to pressure fluctuations. Therefore we did not use an accumulator in the motor supply. (The bearing supply was filtered however.) Rotor shaft vibration was correlated with the motor and bearing pocket pressures and with the supply pressure. See [Fig. 3.2.13]. Typical data are given in [Figs. 3.2.12 and 3.2.13].

3.2.4 Conclusion

The agreement of theory with performance predicts that designs can be realized in hardware with good confidence they will work as desired. Our relatively low sensitivity to power supply fluctuations indicates that with only modest filtering, asynchronous shaft motion will be below levels required for optical quality machining. The decision on the tradeoffs of efficiency and thermal control will be machine dependent but we feel the laminar flow motor should be a candidate in future designs.

REFERENCES

Wils-Moren, W.J., Modjarrad, H., and Read, R.F.J., 1982, "Some Aspects of the Design and Development of a Large High Precision CNC Diamond Turning Machine."*Annals of the CIRP*, Vol.31/1, pp.409-414.

Chen, Chien-Jen, 1985, "A Laminar Flow Motor-Driven Machine Tool Spindle". Stanford University, Stanford, CA, USA, Ph.D. Dissertation.

Kraakman, H.J.J., 1969, "A Precision Lathe with Hydrostatic Bearings and Drive". *Philips Technical Review*, Vol.30, No.5, pp.117-133.

Gijsbers, T.G., 1980, "COLATH, a Numerically Controlled Lathe for Very High Precision". *Philips Technical Review*, Vol.39, No.9, pp.229-244.

Bryan, J.B., 1979, "Design and Construction of an Ultraprecision 84 inch Diamond Turning Machine". *Precision Engineering*, Vol.1, No.13, pp.13-17.

Donaldson, R.R. and Patterson, S.R., 1983, "design and construciton of a Large, Vertical Axis Diamond Turning Machine". *Proceedings of SPIE*, Vol.433, Aug. 24-26, pp.62-67.

McCu, H.K. 1983, "The Motion Control System for the Large Optics Diamond Turning Machine(LODTM)". *Proceedings of SPIE*, Vol.433, Aug. 24-26, pp.68-75.

Barkman, W.E. and Woodard, L.M., 1981, "Upgrading a Production Machine Tool for Precision Maching". Oak Ridge Y-12 Plant, Report Y-2264, Dec.

McKeown, P.A. and Morgan, G.H., 1979, "Epoxy Granite: A Structural Material for Precision Machines". *Precison Engineering* Vol.1, No.4

DeBra, D.B., 1984, "Design of Laminar Flow Restrictors foı Damping Pneumatic Vibration Isolators". *Annals of the CIRP*, Vol.33/1.

DeBra, D.B., 1981, "Damping Vibration for Special Lathe". *American Machinist*, pp.115-116.

Warner, R.H., McCulloch, M. and Howard, J., 1983, "Prototype Pneumatic Isolation and Leveling System for the Ultra-Precision Machine Project". Stanford University Course ME 119.

Modrovsky, J., 1957, *Pump for Viscous Fluids*, US Patent 2.777,394, Jan.

15.

Fazekas, G.A., 1961, *Floating Balanced Doctor Blade or Vane*, 2,969,020, Jan. 24.

Fazekas, G.A. 1961, *Pump for Viscous Liquids*, U.S. Patent 2,992,615, Jul. 18.

Bullough, Q., 1974, *Combined Viscosity Pump and Electric Motor*, U.S. Patent 3,794,447, Feb. 26.

Bardsley, E., 1893, *Rotary Fluid Motor*, U.S. Patent 509,644 , Nov. 28.

Sternlicht, B., 1962, *Pumps*, U.S. Patent 3,037,457, June 5.

Fuller, D.D., 1984, *Theory and Practice of Lubrication for Engineers*, John Wiley and Sons.

Hahn, R.S., 1965, "Some Advantages of Hydrostatic Bearings in Machine Tools". *J. of the American Society of Lubrication Engineers*, Mar.

	Efficiency	Heat in Rotor	Side Forces and Cogging Effects
DC Motor	High	Low	Highest
AC Induction	High	Medium	Medium
Bushless DC	High	Low	Highest
Eddy Current	Low	Highest	Lowest
Definite Motor	Medium	Low	Medium to Low
Hydraulic Motor	High	N/A	Medium
Laminar Flow Motor	Low	N/A	Low (see Chapter 6)

Table 3.2.1 Comparison of motors

Design Parameter	Performance Characteristic (Reference Equation)			
	$T \times \omega$ (3.23)	$H_{fe}(H_{fe})$ (3.57)	$Q_{le}(Q_{le})$ (3.60)	ΔL (3.26)
N	$\frac{1}{2}$	— (1)	— (1)	—
P_{SL}	$\frac{1}{2}$	—	1	1
μ	$\frac{1}{2}$	1	-1	—
R	2	3 (2)	1	—
L	1	— (1)	— (1)	1
l_a	—	1	-1	—
l_s	—	— (1)	— (1)	—
h_g	—	-1	3	—
ω	$\frac{3}{2}$	2	—	—
As an example, the power output, $T \times \omega$, varies with $\mu^{\frac{1}{2}}$ holding the theoretical maximum efficiency constant (Eq. 3.23).				

Table 3.2.2

Design Parameter vs. Performance characteristics

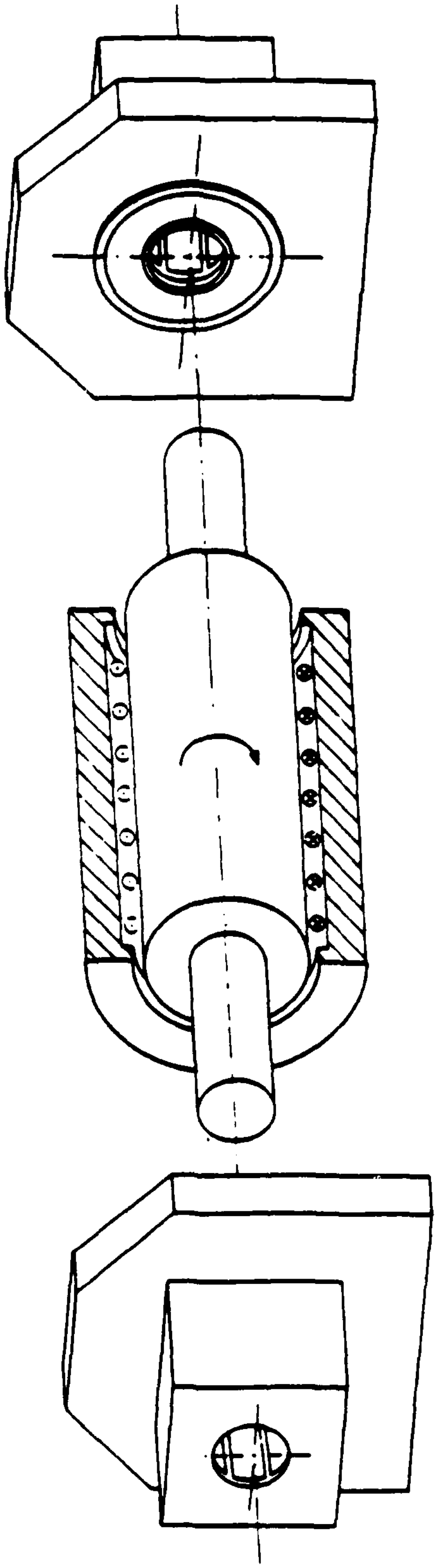


Fig. 3.2.1 Figure of a Laminar Flow Motor and Supporting Hydroelectric Bearings

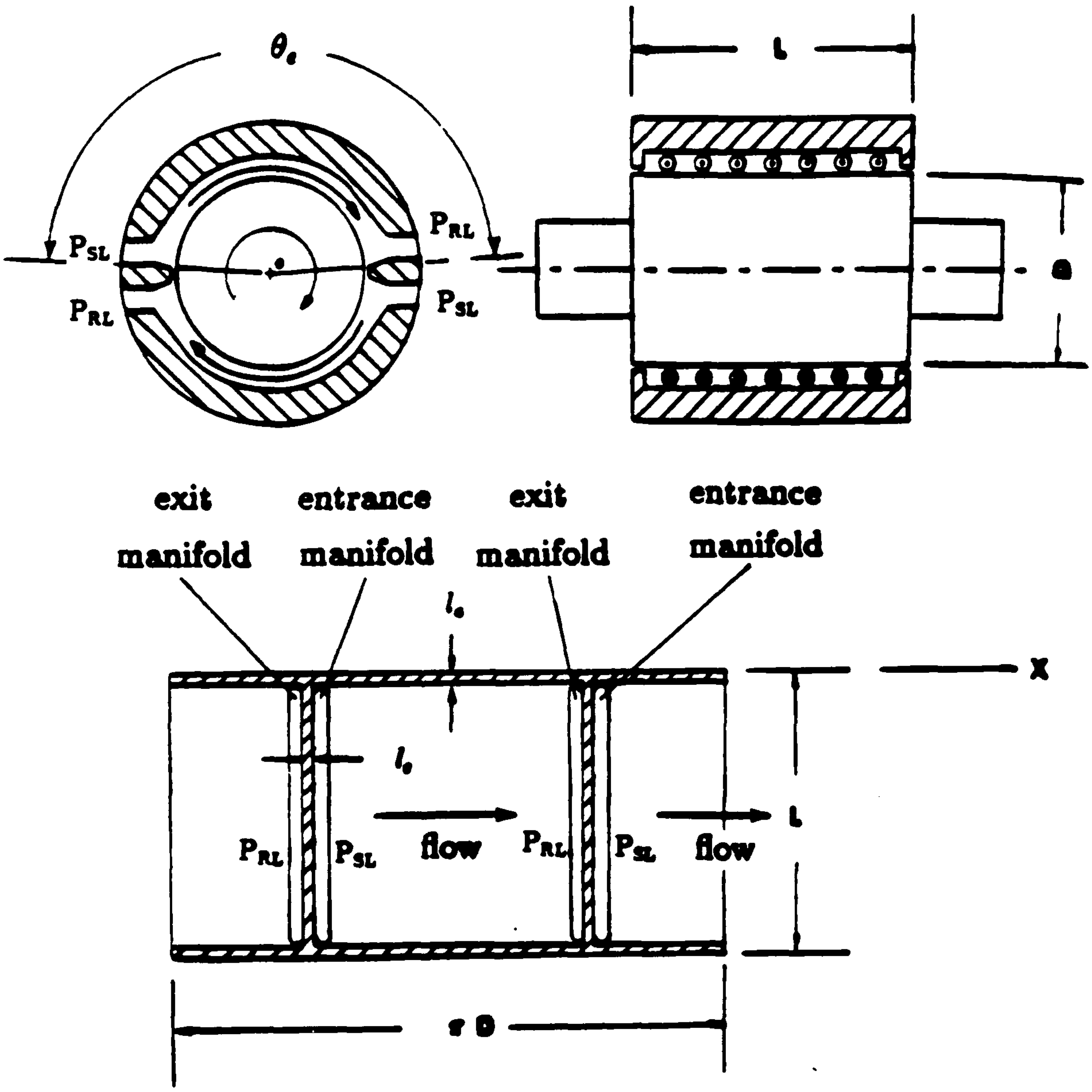


Fig. 3.2.2 Sectional Views of the Laminar Flow Motor

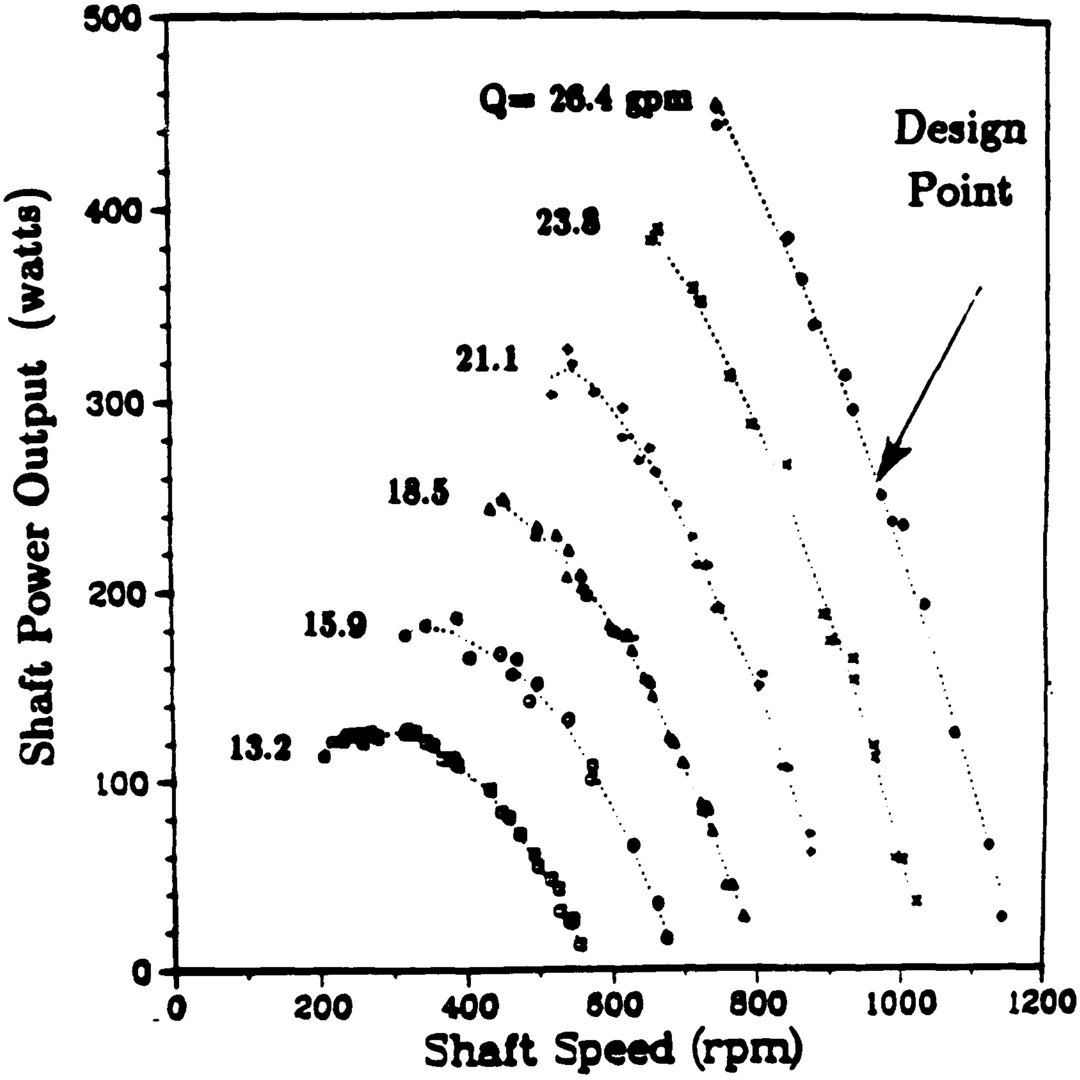


Fig. 3.2.3 Shaft Power output vs. Shaft Speed

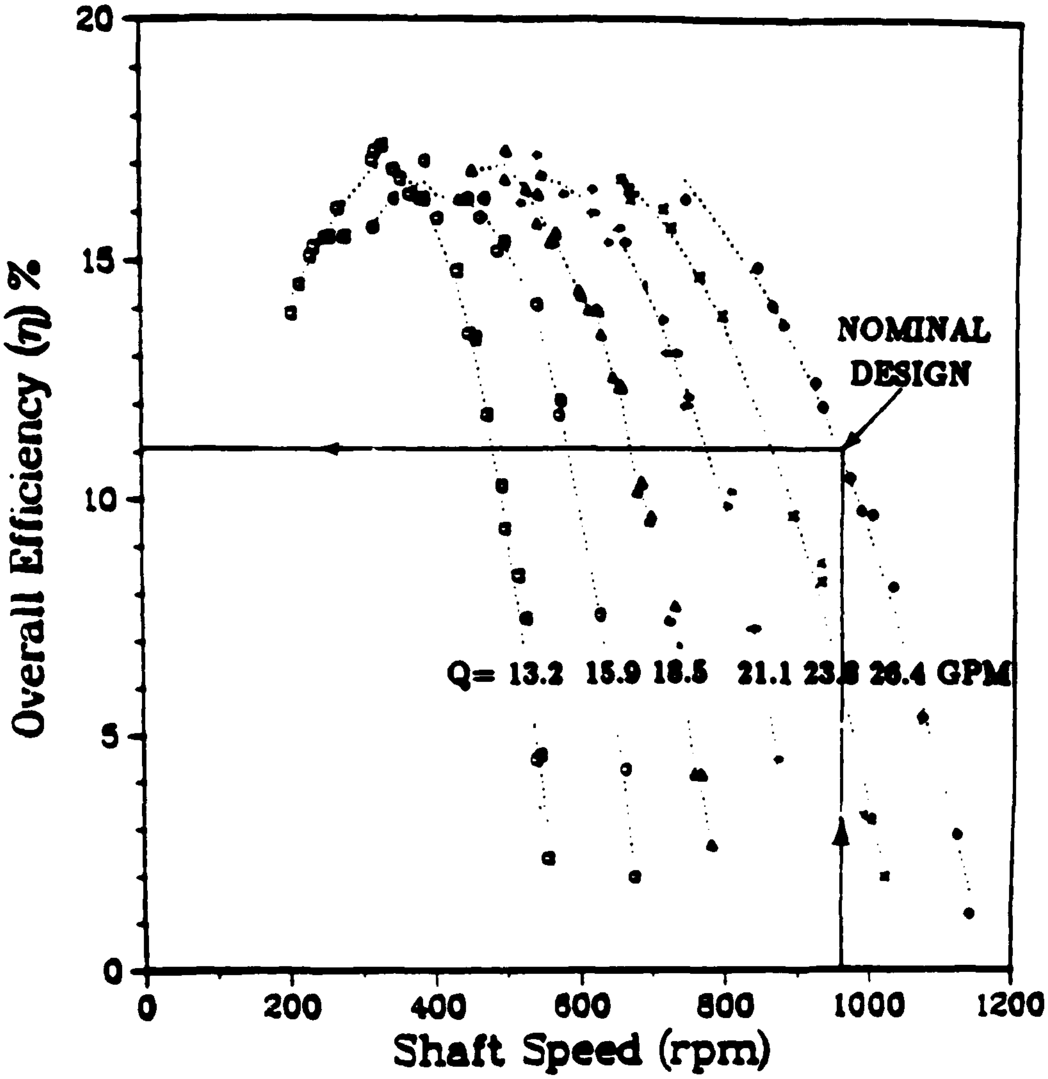


Fig. 3.2.4 Overall Efficiency vs. Shaft Speed

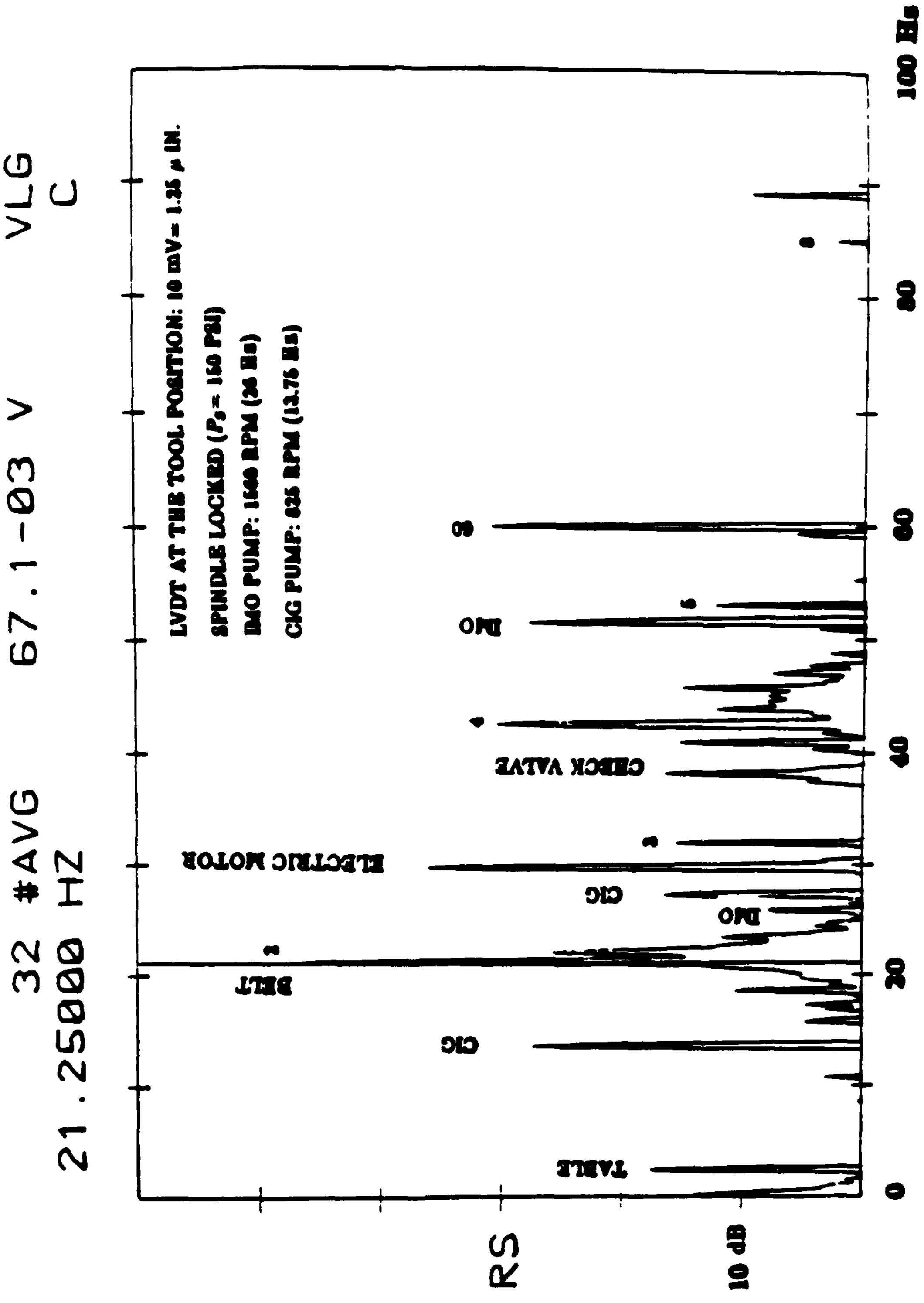


Fig. 3.2.5 Spectrum of Random Motion between Spindle and Diamond Tool

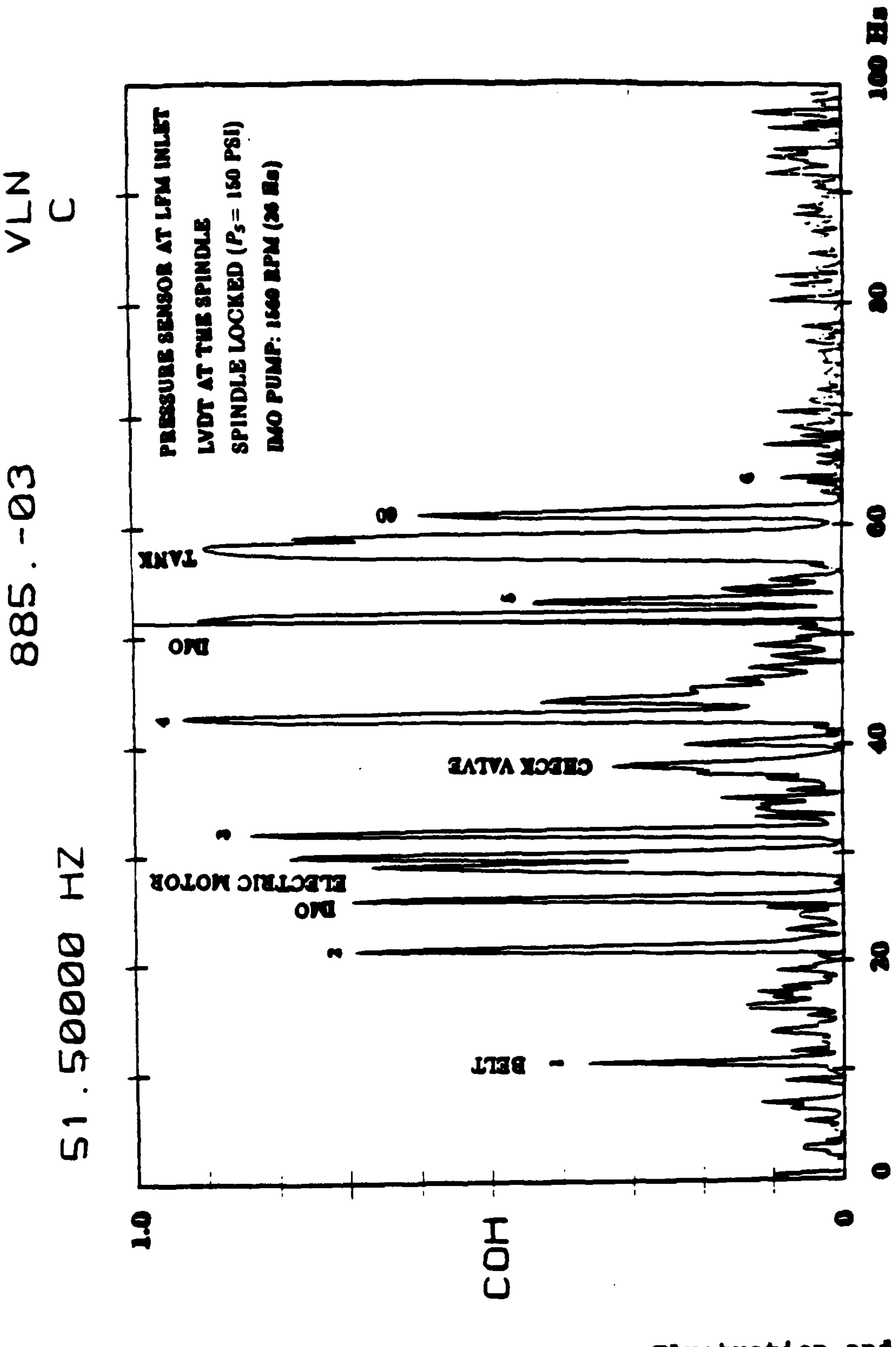


Fig. 3.2.6 Coherence between Pressure Fluctuation and Random Motion

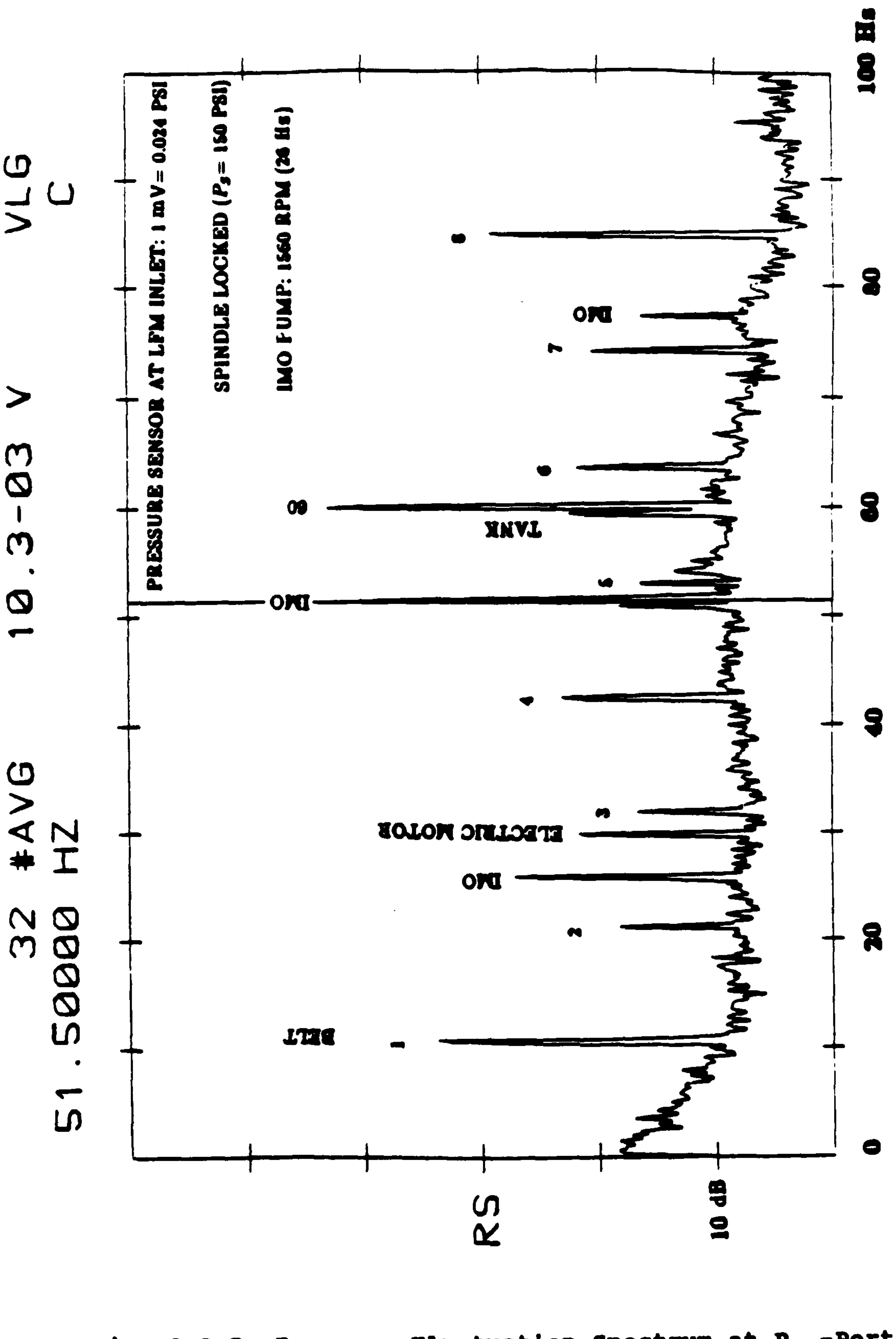


Fig. 3.2.7 Pressure Fluctuation Spectrum at P_{SL}-Port

DETERMINATIONS OF PRESSURE FLUCTUATION

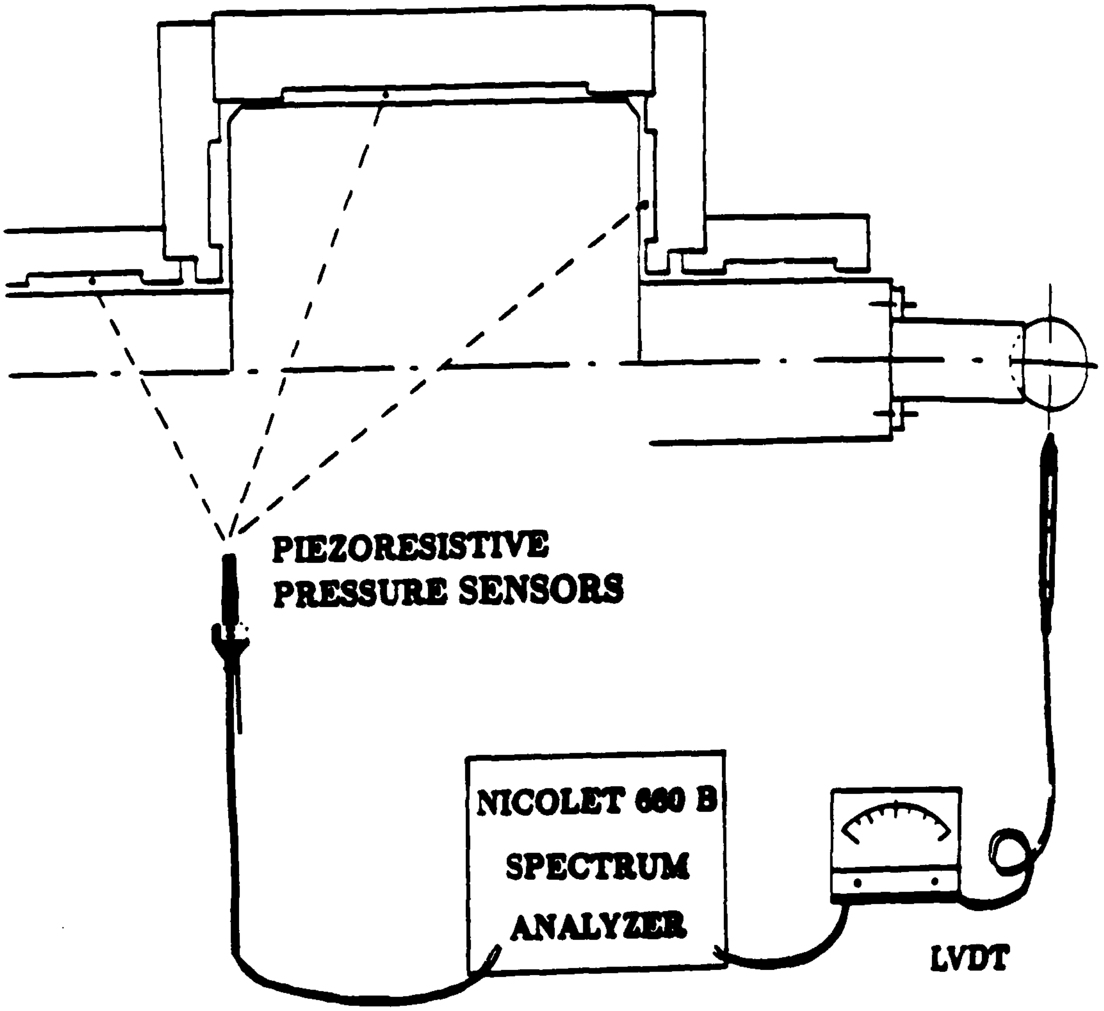


Fig. 3.2.8 Apparatus for Measuring the Smoothness

3.3 DIFFERENTIAL LAMINAR FLOW VALVE

3.3.1 Introduction

The purpose of a differential laminar flow valve is to provide flow to the double acting pistons used to drive each of the stages. The original design was developed with a prototype stage using a cylindrical way and a kinematic design in which the cylinder rode on the piston shaft and was fastened to the stage through a spherical socket for the sixth degree of freedom. This is a convenient design which does not require parallelism between the main cylindrical way and the piston shaft. See Fig. 3.1.6 (in the **ACTUATION** Chapter, *Quiet Hydraulics* Section of the ONR Report). In the final machine tool the ways are a T-cross section which provides the five degrees of freedom needed stiff all in the one construction. The valve for either of these systems has the same requirements. It needs to provide a differential pressure and flow that is free of pressure fluctuations and flow surges.

3.3.2 Selection of a Design

We considered a number of different design options including rotary and translation. Rotary designs can take advantage of available torquers more readily than the linear actuation systems. But we found a solenoid which had adequate stroke and force which could be adapted for our purpose. This involved replacing the sliding bushing with elastic supports which had no stiction. The final design involved two metering disks with laminar flow restriction on each side of each disk. These formed a bridge of resistances which were modulated differentially producing the necessary output pressure and flow. In form this valve is identical to a four way conventional hydraulic servo valve. However, in the hydraulic servo valve, the resistances to flow are modulated orifices. Hence, there is turbulent flow in the conventional design and this turbulence we were concerned, would result in a vibration that would affect the surface finish. Thus we developed the design with the metering disks.

An optimization was carried out for the best radial flow length. Consideration included adequate squeeze film damping so that in the dynamic range of 20 - 100 Hz the valve would appear as a pure integrator. That is the squeeze film damping would be the dominant dynamic effect. The final stroke was 0.1 mm which resulted in adequate radial flow length to gap ratio to insure that the flow was fully developed in the laminar flow restriction.

3.3.3 Experimental Evaluation

The valve and stage arrangement with its piston were completed and experimental evaluation began. At the outset we were surprised to find a static instability which was not expected in the valve which is nominally pressure balanced. However, we found that back pressure in the return flow creates a static instability which increases as the valve departs from its center position. This was compensated for by introducing springs to keep the valve full center. The dynamic tests then showed an unexpected instability which represent two modes at frequencies of 35 and 55 Hz.

The two principal parameters to which the valve is sensitive are the supply pressure and the back pressure for flow leaving the valve.

Tests were started to try to isolate the parameters which were the principal contributors to these unexpected dynamic phenomena. Th valve was modified to provide a flow path to relieve the back pressure. Unfortunately, this made relatively little difference in the behavior.

Subsequent tests have included disconnecting the piston from its support with respect to the base. We have the valve attached to the cylinder which is the moving part of the system. When the actuator moves the stage the valves move with it; hence, there are inertial forces acting on the valve spool. Since this could cause a dynamic feedback, we felt that the disconnecting of the piston with respect to the base would remove the dynamic behavior of the stage. There was relatively little change in the behavior of the system, except a slight reduction in the natural frequencies of the two oscillations. Thus the inertial feedback of itself was not the cause. This test also removed a possible elastic behavior between the piston and the base through its supports which could be one of the natural frequency that was being excited. Obviously, this test disproved this as the principal cause of the dynamic behavior.

In parallel with these experimental test we have increased our simulation efforts.

3.3.4 Simulation

From the beginning we have carried a series of simulations and dynamic analysis of the valve piston combination. The original ones were relatively simple models which were used as the basis for the design. Subsequently, compressibility effects of the fluid, resistance to flow through the channels between the valve and the piston, back resistance to the flow out of the valve, elastic interaction between the piston and the base and a number of other effects have been included. The more complicated twelfth order system is not handled easily by hand and so this work has been reduced to a numerical procedure for evaluation on a program called MATLAB. In parallel with this effort the differential equations were integrated numerically so that nonlinearities could be included. So far none of the simulations of any of the models provide an adequate representation of the dynamic behavior that is being observed.

3.3.5 Conclusion

At this point this portion of research is an unfinished puzzle. We are continuing to work on it since a differential laminar flow valve should be an essential piece of technology available to the designer using quiet hydraulics for precision machine tool control.

3.4 SHORT STROKE HYDRAULIC ACTUATOR

3.4.1 Introduction

We are interested in non-circular cutting on a lathe, specifically, on a diamond-turning lathe. A fundamental limitations of lathes is that they can only be used to manufacture parts which are axisymmetric. Non-circular cutting is a means of making parts which are not bodies of revolution, by actuating the tool in synchrony with the rotation of the spindle.

Diamond-turning can achieve optical quality finishes, since the diamond tool behaves almost as an ideal tool (with little wear, and no built-up edge) when cutting non-ferrous materials, such as copper, aluminum, nickel, or plastics.

What are the applications for non-circular cutting? Consider X-ray telescopes. The mirrors used in X-ray telescopes must be grazing incidence mirrors. In order to get an image, the mirrors are generally paraboloids. The geometric center of the mirror might not be on the optical axis. If we wanted to machine such a mirror on a conventional lathe, we'd have to swing the part about its optical axis, which is unwieldy. An alternative would be to machine the part on its geometric center. However, the part is not axisymmetric about its geometric center. This is where non-circular cutting is useful.

Let's take a paraboloid of revolution. Suppose we wanted to make just a segment of it, off of the optical axis (this is an off-axis paraboloid (see Figure 3.1)). The equation describing a paraboloid of revolution is

$$z = c(x^2 + y^2) \tag{3.5}$$

When the paraboloid of revolution is translated and rotated to a different coordinate system, the result gets complicated. We get

$$\begin{aligned} 0 \;=\; & (A_{11}x' + A_{21}y' + A_{31}z' + x_0)^2 \qquad (3.6)\\ & + (A_{12}x' + A_{22}y' + A_{32}z' + y_0)^2 - (A_{13}x' + A_{23}y' + A_{33}z' + z_0) \end{aligned}$$

where the A_{ij}'s are the direction cosines. In general, we have reference frame xyz (the optical axis is z), and reference frame $x'y'z'$ (with origin x_0, y_0, z_0), the frame whose origin is the geometric center of our off-axis paraboloid. The direction cosine matrix A is found such that $AA^T = I$; $|A| = 1$. Then, $\vec{x'} = A(\vec{x} - \vec{x_0})$; $\vec{x} = A^T\vec{x'} + \vec{x_0}$.

Once the surface has been described, it may then be transformed to cylindrical coordinates. Then, the shape of the off-axis paraboloid can be expressed as a Fourier series in terms of the polar angle θ and the distance r from the geometric center.

A simpler surface, which is not axisymmetric, is an inclined plane. Take a plane, and have its normal tilted with respect to the spindle axis. This is obviously no longer a figure of revolution with respect to the spindle axis; metrology to measure figure and finis.ı is simple, and in the frame of reference of the spindle, the figure of the tilted plane is given by a sinusoid.

3.4.2 Actuation

Since we are interested in non-circular cutting, it is necessary to have a translational actuator with a response similar in speed to the spindle rotation.

Means of Actuation

Let us consider what types of translational actuators can be used. Direct-acting mechanisms can be used, such as solenoids or hydraulic pistons, or an indirect mechanism, such as a rotary motor driving a lead-screw. In addition, the actuator itself can be electromagnetic (such as conventional motors), pneumatic, hydraulic, or piezoelectric.

What are the design requirements for our translational actuator? First, we are only interested in a very small range of motion, i.e. a throw of less than 1 mm. Second, we would like fast action from the actuator. Third, since we want a good finish, the actuator must be stiff. An actuator with linear dynamics would also make the issue of controlling the system easier. Similarly, long term stability of the actuator parameters are desirable.

Piezoelectric actuators generally can provide high bandwidth, but their range of motion is limited. An electromagnetic actuator, such as a motor driven lead-screw, can have a large range of motion. For instance, the ways of an engine lathe are driven with lead-screws. Likewise, the head of a floppy disk drive is driven with a lead-screw. Lead-screw mechanisms, however, tend to be slow. Direct acting electromagnetic mechanisms, such as the voice coil actuator used in disk drives, have both a fast response and a fairly large range of motion, but are not very stiff.

Higuchi [7] has demonstrated the manufacture of non-axisymmetric parts on a lathe, which he calls "non-circular cutting." He uses conventional hydraulic servos and a conventional lathe; turning parts at 10 to 100 rpm. Patterson developed a piezoelectric tool-post actuator at Lawrence Livermore National Laboratories [9]. This tool-post servo has a range of 2.5 μm and a bandwidth of 100 Hz. Dow, at North Carolina State University, also developed a piezoelectric tool-post actuator [8],[3]. This actuator has a range of 20 μm, and a bandwidth of approximately 1000 Hz.

An ideal actuator would have high bandwidth, stiffness, and a large range of motion. In addition, this actuator would be insensitive to environmental disturbances, such as temperature changes, and would generate little disturbance. An actuator possessing all these properties, unfortunately, does not exist.

We chose to use hydraulic actuation for the short stroke tool actuator. The choice of hydraulics was made because a large amount of power is available in a small package; the working fluid also carries away heat generated by inefficiencies, and the choice of hydraulic actuation complements the laminar flow devices developed in this project [2].

Design of the Actuator

Reiterating the desireable characteristics of an actuator, we would like high bandwidth, stiffness, linearity, and good disturbance rejection. Where the actuator's performance is weak, feedback may be used to improve its behavior.

A problem frequently encountered with hydraulic actuators is due to the compressibility of the working fluid. The resonance due to the compressibility of the fluid can be altered by changing the volume of the actuator. Resonances may also occur in the pipes.

The short stroke actuator consists of a diaphragm constrained by a flexure. The pressure in the diaphragm is controlled by a flapper valve (in a pressure bridge arrangement); the flexure provides both a safe enclosure in case the diaphragm breaks, and stiffness in all but one degree of freedom.

Figure 3.2 shows the construction of the actuator, and figure 3.3 shows the construction of the diaphragm. When pressure is applied to the diaphragm, plate stresses develop. Because of symmetry, this situation can be treated as a two-dimensional stress problem. If we assume that the diaphragm is linear and isotropic, the stresses and strains can be found by elasticity theory (for instance, see Timoshenko [12]).

When body forces are absent, we can find an exact solution for the stresses and strains in the diaphragm by solving the biharmonic equation, $\nabla^2(\nabla^2\phi) = 0$ (this is also written as $\nabla^4\phi = 0$). ϕ is a stress function, from which the principal stresses and the shears may be derived.

We will assume axisymmetric stress and deformation. It is more convenient to express the stresses in the diaphragm in terms of cylindrical coordinates. Thus, $\nabla^4\phi = 0$ becomes

$$\left[\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2}{\partial\theta^2} + \frac{\partial^2}{\partial z^2}\right]\left(\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial\phi}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2\phi}{\partial\theta^2} + \frac{\partial^2\phi}{\partial z^2}\right) = 0. \quad (3.7)$$

With ν being Poisson's ration, the principal stresses are given by

$$\begin{aligned}
\sigma_r &= \frac{\partial}{\partial z}\left(\nu\nabla^2\phi - \frac{\partial^2\phi}{\partial r^2}\right) \\
\sigma_\theta &= \frac{\partial}{\partial z}\left(\nu\nabla^2\phi - \frac{1}{r}\frac{\partial\phi}{\partial r}\right) \\
\sigma_z &= \frac{\partial}{\partial z}\left((2-\nu)\nabla^2\phi - \frac{\partial^2\phi}{\partial z^2}\right) \qquad (3.8) \\
\tau_{rz} &= \frac{\partial}{\partial r}\left((1-\nu)\nabla^2\phi - \frac{\partial^2\phi}{\partial z^2}\right)
\end{aligned}$$

The stresses in the diaphragm depend on r and z; they do not depend on θ. Moreover, two of the six stress components, $\tau_{r\theta}$ and $\tau_{\theta z}$ are zero. A function ϕ which satisfies the biharmonic and the boundary conditions will yield the stresses and strains in a plate.

The boundary conditions are determined from the loading on the plate, and the plate geometry. At the edge of the plate, we have a cantilever condition. At the inner radius of the plate, where the thickness is increased, a cantilever condition is a good approximation (since the moment of inertia of the structure is proportional to the cube of the thickness). Finally, the loading on the plate is uniform pressure. Thus, at $z = -t/2$, $\sigma_z = P$, the pressure of the working fluid. Likewise, at $z = t/2$, $\sigma_z = 0$.

From axisymmetry, equation 3.7 becomes

$$\left[\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\right) + \frac{\partial^2}{\partial z^2}\right]\left(\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial \phi}{\partial r}\right) + \frac{\partial^2 \phi}{\partial z^2}\right) = 0. \tag{3.9}$$

Note that if ϕ satisfies $\nabla^2\phi = 0$, it will also satisfy $\nabla^4\phi = 0$.

We start with the Laplacian

$$\left(\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial \phi}{\partial r}\right) + \frac{\partial^2 \phi}{\partial z^2}\right) = 0. \tag{3.10}$$

This is a partial differential equation in r and z. Applying the following changes of variables (to spherical coordinates R and ψ),

$$\begin{aligned} z &= R\cos\psi \\ r &= R\sin\psi \\ x &= \cos\psi \\ R &= \sqrt{r^2+z^2} \end{aligned} \tag{3.11}$$

we get

$$\frac{\partial^2\phi}{\partial R^2} + \frac{2}{R}\frac{\partial\phi}{\partial R} + \frac{\cos\psi}{R^2\sin\psi}\frac{\partial\phi}{\partial\psi} + \frac{1}{R^2}\frac{\partial^2\phi}{\partial\psi^2} = 0. \tag{3.12}$$

If we assume solutions of the form $\phi = \sum_{n=0}^{\infty} a_n R^n \psi_n$, (a_n's are arbitrary constants) we can use separation of variables. We obtain an ordinary, second order, differential equation

$$(1-x^2)\frac{d^2\psi}{dx^2} - 2x\frac{d\psi}{dx} + n(n+1)\psi = 0. \tag{3.13}$$

Equation 3.13 is Legendre's differential equation. The solutions for Legendre's equation are Legendre polynomials. If we let $\psi = P_n(x)$, where $P_n(x)$ is an nth order Legendre polynomial (in x), then

$$\phi_n = a_n R^n P_n \tag{3.14}$$

is a solution to equation 3.12. We can return to cylindrical coordinates by applying the transformation of equation 3.11.

A particular solution to equation 3.9 can be formed by

$$\phi_n = b_n R^{n+2} P_n(x) \tag{3.15}$$

(b_n's are constants). Appropriate solutions to the biharmonic can be found by taking linear combinations of $R^n P_n(x)$ and $R^{n+2} P_n(x)$, then, solving for the a_n's and b_n's that match the boundary conditions.

Consider $\phi = a_0 P_0(x) + \ldots + a_6 R^6 P_6(x) + b_0 R^2 P_0(x) + \ldots + b_6 R^8 P_6(x)$. An approximate solution can be obtained by guessing

$$\phi = A_0 + B_0 r^2 + C_0 \ln r + D_0 r^2 \ln r + \mathcal{F}_0(r) \tag{3.16}$$

and matching the coefficients to the boundary conditions.

By superposition of tabulated plate functions (see Roark and Young [10]), we can get an approximation for the deflection at the centerline of the diaphragm as a function of the internal pressure. The deflection z is

$$\begin{aligned} z &= \frac{6Pa^5(1-\nu^2)}{Ebt^3}\left(\frac{(1-(\frac{b}{a})^2(1+2\ln\frac{a}{b}))\frac{b}{a}((\frac{b}{a})^2-1+2\ln\frac{a}{b})}{8(1-(\frac{b}{a})^2)}\right. \\ &\quad \left. -\frac{\nu}{4a}(((\frac{b}{a})^2+1)\ln\frac{a}{b}+(\frac{b}{a})^2-1)\right) \end{aligned} \tag{3.17}$$

Since E is Young's modulus; ν is Poisson's ratio; and a, b, and t are geometry constants, then z is a linear function of the pressure P.

Hydraulic Design

To cause tool movement, the diaphragm must be pressurized. In order to modulate the tool movemer', you need to control the pressure inside the diaphragm. The pressure is controlled via the flapper valve. What are the dynamics associated with the flapper valve, piping, diaphragm, and flexure?

When the tool is moved, there is small change of volume inside the diaphragm. There must be oil flow, either into the diaphragm, or out of it. In addition to viscous losses inside the diaphragm and its feed tube, there may be effects from the compressibility and 7inertance of the oil. Finally, the tool post and flexure contribute inertia to the system.

First, we can calculate the viscous losses in the feed tube. If the actuator is excited with a sinusoid, we know that the volume change in the diaphragm reaches a maximum within one quarter of a cycle. We assume fully developed flow, and assume a triangle wave as an approximation for a sine wave (i.e. linearize the sine wave).

The pressure loss for fully developed laminar flow in a pipe is given by

$$\Delta P = \frac{Q128\mu L}{\pi D^4} \tag{3.18}$$

where Q is the flow rate; μ is the viscosity of the fluid; L is the pipe length, and D is the pipe diameter. The flow rate is estimated by dividing the volume change in the diaphragm by one quarter of the period of the excitation.

Pipeline losses (due to the fluid compressibility and inertance) may also be calculated. Viersma [17] or Wylie and Streeter [18] show methods for calculating pipeline resonances.

The pressure distribution along a pipe may be expressed by the wave equation

$$\frac{\partial^2 p}{\partial x^2} = CL\frac{\partial^2 p}{\partial t^2} + RC\frac{\partial p}{\partial t} \tag{3.19}$$

with R, C, and L being fluid and pipe resistance, capacitance, and inertance. Solutions to this equation are obtained by assuming $p = X(x)T(t)$, and using separation of variables.

For our geometry (pipe length of 40 mm; diameter 4 mm; oil specific gravity 0.9; viscosity 0.1 Pa-sec; compressibility 0.8 GPa), the pipeline losses are less than 5% at 150 rad/sec.

Cross Slide Flexure

For useful work, the tool must move in two axes. The fast hydraulic actuator moves parallel with the spindle axis. We need a second movement for a cross slide.

The cross slide needs a longer stroke, but need not be as fast. Using a flexure to eliminate the problems of coulomb friction, a cross slide was built. The flexure is based on stacking two parallelogram flexures (see Figure 3.4). Each parallelogram contributes an error perpendicular to the main motion, but by reversing one parallelogram on top of the other, these errors cancel. The flexure design is compliant in the direction of travel for the cross slide, and stiff in all other axes.

3.4.3 Sensors

What is optical quality finish? It's a surface finish condition where the reflections are specular, not diffuse. This is generally obtained when the peak-to-valley roughness of the surface is about a tenth of the wavelength of light. Thus, we need peak-to-valley roughnesses of 50 to 75 nm (or about 3 microinches).

What actuators and sensors are needed? We need to actuate the spindle and sense the spindle angle (theta); we need to actuate the tool in the x-direction (cross-slide) and sense x-position (x), and we need to actuate and sense the tool in the z-direction. However, we need only closed-loop feedback in z (since that is the high bandwidth part). Suppose we wanted to turn our part at about 1000 rpm; this is 16 revolutions/second. To get a tilted plane, we need

$$z = kx\sin(\theta + \theta_0) \tag{3.20}$$

Obviously, if we had no errors, we'd get a perfect finish. In fact, we can determine from theoretical finish considerations the maximum errors allowable in θ, x, and z. These are "repeatability" type maximum errors, i.e. systematic errors only affect figure, not finish. From geometry and other theoretical finish arguments, we calculate that the desired repeatability in θ is 0.0005 radians, repeatability in x is 20 microns, and repeatability in z must be better than 50 nm.

LVDT's

A variety of transducers can be used to measure displacement. We have two different needs for a translation sensor: One sensor is needed to measure the cross-slide movement, and another is required to measure the tool post movement.

The hydraulic cross slide has a travel of approximately 12 cm (5 in); the flexure based cross slide has a travel of approximately 5 mm (0.2 in). From theoretical finish considerations, assuming a tool radius of 1 mm (0.040 in), we need to control the feed to within 0.020 mm (0.0008 in). Thus, the cross-slide sensor has to resolve to one part in 10^4. A Linear Variable Differential Transformer (LVDT) is used as the cross-slide sensor.

Similarly, the tool post actuator has a total travel of 0.25 mm (0.010 in). From theoretical finish calculations, we need to control the tool position to within 75 nm (3 μin). Again, the resolution required is approximately one in 10^4. An LVDT is also used for the tool post actuator sensor.

Since we need optical quality finish and figure, it would be natural to assume that we would want to use laser interferometers. Laser interferometers can provide excellent dynamic range. Unfortunately, laser interferometers are susceptible to contamination, and they are rather expensive.

LVDT's are rugged; oil contamination is not a problem for these devices. They are also inexpensive, and since their output is analog in nature, they lend themselves readily to analog feedback systems. Commercial signal conditioning equipment for LVDT's are cumbersome. Their bandwidth is low, and they are designed to operate a variety of LVDT's. We have designed a circuit for driving LVDT's. The components on the circuit are picked to match the characteristics of each LVDT used, optimizing the excitation frequency. Similarly, a six pole filter (implemented as a cascade of 3 2-pole filters) is used to filter the output. Sidman [11] tested this circuit with a RVDT (LVDT for sensing angular displacements), and found repeatability better than 5×10^{-5}.

Rotary encoder

Since we are interested in making non-axisymmetric parts, we also need a sensor for the spindle angle. Two different types of transducers can be used to measure rotation angle: Resolvers and encoders. Although tachometers are frequently used in motor speed control applications, tachometers measure only speed. Resolvers provide an analog output which is the sine of the angle being measured (the quadrature output of the resolver gives the cosine). Since resolvers are analog sensors, the resolution of a resolver is limited to the quality of the electronics package. Encoders have a resolution limited to the number of lines. Linear optical encoders are available with up to 10,000 lines/inch.

Design constraints force us to use an encoder to determine spindle angle. In addition, since the working environment is oily, we use a magnetic pickup in the encoder. Thus, the rotary encoder is a carbon steel disk, with a series of holes cut into it. A Hall effect sensor, along with an alnico permanent magnet, senses the presence (or absence) of a hole. The sensor output can be converted to digital pulses coinciding with the signal peak.

Consider the figure we want to fabricate: We express the figure in cylindrical coordinates as $z = Kx\sin(\theta + \theta_0)$. How much error in θ would result in an unacceptable error in z? We need θ repeatability of 0.0005 radians.

Since our encoder has only 32 holes, one would expect to obtain, at best, a resolution of 0.2 radians. By using other available information, we could get better resolution from this encoder.

Suppose, for instance, that the encoder was used to provide feedback for speed control of the spindle. Since the spindle speed is now regulated, it is simple to integrate the spindle velocity to obtain the spindle angle. The accuracy of the spindle angle measurement would no longer depend directly on the number of pulses available from the encoder; it would now depend on the accuracy of the speed regulation.

The spindle is driven by flow; a certain amount of flow through the spindle generates shear stresses [2]; these stresses can be converted to a driving torque. This driving torque is resisted by the inertia of the spindle, the viscous friction of the spindle bearings, and disturbance torques, such as cutting forces. We can write equations of motion for the spindle:

$$J\ddot{\theta} + B\dot{\theta} + T_w - T_u = 0 \tag{3.21}$$

where θ is the spindle angle, J is the moment of inertia of the spindle, B is the viscous friction, T_w is the total of the disturbance torques, and T_u is the driving torque.

Even though the spindle angle sensor, by itself, is incapable of resolving better than 0.2 radians, we should be able to estimate the spindle angle to a much better resolution. How accurately can the spindle angle be estimated? And, just as important, how do you mechanize the estimator for the spindle angle?

Consider the spindle estimator as a Kalman filter. We form a state vector $x = [\theta\, \dot{\theta}]'$. Then, Equation 3.21 can be rewritten as

$$\dot{x} \equiv \begin{pmatrix} \dot{\theta} \\ \dot{\theta} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & -\frac{B}{J} \end{pmatrix} \begin{pmatrix} \theta \\ \dot{\theta} \end{pmatrix} + \begin{pmatrix} 0 \\ -\frac{1}{J} \end{pmatrix} (T_w - T_u) \tag{3.22}$$

or,

$$\dot{x} = Ax + Bu + B_1 w \tag{3.23}$$

The estimator state vector is $\hat{x} = [\hat{\theta}\ \dot{\hat{\theta}}]$, where $\hat{\theta}$ and $\dot{\hat{\theta}}$ are the estimated spindle angle and velocity. Again, we write

$$\dot{\hat{x}} = A_e\hat{x} + B_e u + B_{e_1} w. \tag{3.24}$$

To ensure that the difference between the estimator states and the actual spindle states are small, the spindle measurements are fed back to the spindle estimator. Suppose the measurement y is available, where $y = Cx$. Then, the state estimator becomes

$$\dot{\hat{x}} = A_e\hat{x} + B_e u + L(y - C_e\hat{x}). \tag{3.25}$$

Here, L is the estimator gain. For a Kalman filter, the design of L is based on the noise characteristics of the system.

Since the available angle measurements are discrete (we have an encoder), a discrete estimator design is indicated. Now, instead of $\hat{x}(t)$, we have $\hat{x}(k)$, with the estimates (and the measurements) only available as discrete sequences.

We need angular information to within 0.0005 radians; our encoder provides a resolution of 0.2 radians. Therefore, interpolating 400 estimates between measurements will provide us the necessary resolution. Consider, then, the following estimator formulation:

1. Make a measurement.
2. Do a measurement update on the estimator.
3. Do a time update on the estimator (i.e. run the estimator open loop).
4. Repeat the time update 400 times.
5. Go back to *item 1.*

In fact, the implementation of such an estimator yields poor performance. Hashemi and Laub [6] investigated this architecture for faster computation, but the results are just as valid when this filtering scheme is used as an interpolator.

A much better interpolator may be used by implementing a phase-locked loop on the encoder.

3.4.4 Controls

Given perfect measurements of θ, x, and z, and assuming that $\dot{\theta}$ is constant, how do you control z? And how do you make this control method useful for general purposes?

You already know what kind of shape you want to make before you make it. So instead of commanding the tool to do something, you pose the problem as that of following a path. This is where

Feedforward

comes in. In feedback control, you take the measured output, and compare that with the desired output. You use the *error* to control your plant. Suppose you want to have your plant follow a quickly changing signal. If you only use the error signal for control, you're going to need high feedback gains; this can lead to stability problems. With feedforward, you look at the desired signal, and if it changes, you immediately change the plant's input. Feedback is then used to take care of the residual errors (see Figure 3.5).

We explore the following design methods: Feedforward augmentation; design of feedforward via plant inverse; incorporation of a model of the input dynamics, and use linear quadratic synthesis. [15]

The surfaces we will machine can be expressed in terms of Fourier series; therefore, we wish to investigate the response of the controller and plant to a periodic input. Furthermore, there are many possible error sources: The plant model will not be an exact representation of the plant; sensor and actuator noise are present; and the input signal will vary both in frequency and amplitude.

Model Description

Our hydraulic actuator (and its associated drivers and sensors) can be modelled as a second order damped oscillator. The pole locations are at $(-120 \pm 120i)$ *radians/second*. For convenience in the simulation runs we will present, we have scaled time by 10, giving normalized model roots of $(12 \pm 12i)$. We will use a sampling rate of 500 Hz, scaled to 50 for simulation. We want to track sinusoids around 100 radians/second, scaled to 10.

The plant model can be written as

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k \\ y_{k+1} &= Cx_{k+1} \end{aligned} \tag{3.26}$$

With the continuous domain poles at $(-12 \pm 12i)$, the sampled system is

$$\begin{aligned} x_{k+1} &= \begin{bmatrix} 0.577 & -4.49 \\ 0.016 & 0.951 \end{bmatrix} x_k + \begin{bmatrix} 0.016 \\ .0002 \end{bmatrix} u_k \\ y_{k+1} &= \begin{bmatrix} 1.0 & 15.0 \end{bmatrix} x_{k+1} \end{aligned} \tag{3.27}$$

In transfer function form, this is

$$H_p(z) = \frac{0.0181z - 0.0134}{z^2 - 1.2562z + 0.4176} \tag{3.28}$$

We want the plant to follow the reference signal $r = \sin 10t$. The simulations to study controller performance will reflect that the amplitude and frequency of the reference may vary, that there may be input noise, and that the plant model may not be a perfect match for the actual plant.

Feedforward Augmentation

Classical design methods for feedforward involve first, designing the feedback controller. The feedback controller uses the difference beteen the output y and the command input r. The control signal is

$$u = -K(y - r) \tag{3.29}$$

where K is the controller.

Minimizing transient errors requires high gains. The plant model used in designing the controller is only an approximation to the real plant. For instance, an anti-aliasing filter in the plant sensors may not be modelled. The presence of A/D and D/A converters adds quantization errors; these may be treated as a noise source, whose effects are increased with gain. High gains can lead to stability problems of the feedback loop.

A feedforward block may be added, so that changes in the command input above the feedback bandwidth are fed directly to the plant, allowing the feedback controller to take care of lower frequencies. The control signal becomes

$$u = -K(y - r) + Fr \tag{3.30}$$

with F being the feedforward controller. In classical designs, the feedback controller is designed first, and the feedforward is generally designed as a bandpass filter times the plant inverse to enhance the bandwidth of the system.

Since the poles and zeros of the model in Eq. 3.28 are well inside the unit circle, a complicated feedback compensator appears to be unnecessary. Assume a controller of the form given in Eq. 3.30. A feedforward controller $F(z)$ designed to give the desired response would require $y/r = 1$, where

$$\frac{y}{r} = \frac{HF + HK}{1 + HK} \tag{3.31}$$

from which $F = 1/H$. This is not practical to mechanize over all frequencies of interest, nor is it required. To extend the bandwidth of the system, we only need F to approximate $1/H$ in the range from the feedback bandwidth to the bandwidth required.

We can design $K(z)$ independently of $F(z)$; for instance, a gain of 3 would put the system poles at $0.61 \pm 0.13i$, giving a 1% settling time of $10\,T$.

A simpler feedforward design is possible when the command is restricted to a narrow band of frequencies. For example, we could use

$$F(z) = \frac{z^2 - 0.9825z}{2(z - 0.8187)^2} \tag{3.32}$$

This is a bandpass filter at a scaled frequency of 10.

The results of this design and the other designs will be compared later.

Feedforward by Inverse

It is tempting to design the feedforward block to be the inverse of the plant dynamics, giving a total transfer function of 1 (Tomizuka and others, 1985, 1986).[13] [14] This approach is sensitive to errors in modelling, and can lead to instability if the plant has non-minimum phase zeros. We can separate the plant numerator into parts inside and outside the unit circle. Thus, we take the closed loop transfer function, and partition it into

$$H(z) = \frac{N_s(z)N_u(z)}{D(z)} \tag{3.33}$$

where $N_s(z)$ are the closed loop zeros inside the unit circle, and $N_u(z)$ are the non-minimum phase zeros.

The feedforward controller is designed as a command preprocessor, with the structure

$$F(z) = \frac{D(z)}{N_s(z)N_u^*(z)} \tag{3.34}$$

The choice of $N_u^*(z)$ is made to minimize the error between r and y. In this case, they suggest using

$$N_u^*(z) = \frac{|N_u(1)|^2}{N_u(z)} \tag{3.35}$$

The resulting overall transfer function is

$$H(z) = \frac{D(z)N_u(z)}{N_s(z)|N_u(1)|^2} \frac{N_s(z)N_u(z)}{D(z)} \tag{3.36}$$

Error Space

Another approach is to model the reference input r. By modelling the input, and incorporating the system error $y - r$ as states, we get the equivalent of integral control, avoiding open loop feedforward and obtaining better robustness (Franklin and Emami, 1982). [4]

First, we need to augment the state space model by incorporating error states. Since we are interested in tracking sinusoids, we can describe r as a second order differential equation (in fact, the order of r increases by 2 for each harmonic that we wish to track). For our case (following sin 10t with a sampling rate of 50), we express r as a difference equation

$$r_{k+1} = 1.96r_k - r_{k-1} \tag{3.37}$$

Since the command input satisfies a second order equation, we create two error states, the first being

$$e_k = y_k - r_k \tag{3.38}$$

and the other, e_{k+1}.

An augmented state vector is formed from the original states and the error states. If we define the augmented state vector (and its dynamics) by

$$\begin{aligned} \psi_{k+1} &\equiv A_e\psi_k + B_e\mu_k \\ \psi_k &\equiv [\, e_k \;\, e_{k+1} \;\, \xi_k \;\, \xi_{k+1} \,]' \end{aligned} \tag{3.39}$$

where ξ is related to x by the dynamics of the command (Eq. 3.37), i.e.

$$\xi_k \equiv x_{k+2} - 1.96x_{k+1} + x_k \tag{3.40}$$

and μ is a similar transformation on the control signal u:

$$\mu_k \equiv u_{k+2} - 1.96u_{k+1} + u_k \tag{3.41}$$

A_e (Eq. 3.39) and B_e are, for our case,

$$A_e = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 1.96 & [C] & \\ 0 & 0 & & \\ 0 & 0 & & [A] \end{bmatrix} \tag{3.42}$$

and

$$B_e = [\, 0 \;\, 0 \;\, B' \,]' \tag{3.43}$$

The feedback gains of $\mu = -K_e\psi$ can be computed by linear quadratic synthesis or pole placement methods to give satisfactory dynamics to the system in Eq. 3.39. We then expand the feedback to

$$\begin{aligned} \mu_k \;\; = \;\; & -K_e(1)e_k - K_e(2)e_{k+1} \\ & - K_e(3)\xi_k - K_e(4)\xi_{k+1} \end{aligned} \tag{3.44}$$

and transform μ back to u and ξ back to x (Eq. 3.41 and Eq. 3.40) to get the control law. The control signal becomes

$$\begin{aligned} u_k = & -K_e(3,4)x_k \\ & + 1.96(u_{k-1} + K_e(3,4)x_{k-1}) - K_e(2)e_{k-1} \\ & - (u_{k-2} + K_e(3,4)x_{k-2}) - K_e(1)e_{k-2} \end{aligned} \tag{3.45}$$

Model-following Feedforward

Another design method involves using the states of the model of the input explicitly. We wish the plant output to track an input r which satisfies

$$\begin{aligned} x_{r,k+1} &= A_r x_{r,k} \\ r_k &= C_r x_{r,k} \end{aligned} \tag{3.46}$$

Note that this is just Eq. 3.37 in state space form. By requiring $(y - r) \rightarrow 0$, we can design feedforward gains based on the states of the input model x_r and feedback gains based on the plant states x.

The compensation is of the form

$$u = -Kx + \mathcal{F}(x_r) \tag{3.47}$$

where K may be chosen by pole placement methods or linear quadratic synthesis. The control law of Eq. 3.47 requires using the states of the model for the reference input, although usually, only the reference input itself is accessible. From Eq. 3.46, we can build an estimator for $x_{r,k}$:

$$\hat{x}_{r,k+1} = A_r\hat{x}_{r,k} + L(r_k - C_r\hat{x}_{r,k}) \tag{3.48}$$

L being the reference estimator gain. The estimated reference states are then used for feedforward.

Trankle and Bryson (1978) [16] show that writing Eq. 3.47 as

$$u_k = -Kx_k + (U + KX)x_{r,k} \tag{3.49}$$

and choosing U and X (the feedforward gains) to solve the system

$$\begin{aligned} CX &= C_r \\ AX - XA_r + BU &= 0 \end{aligned} \tag{3.50}$$

will make the error $y - r$ go to zero. Although Eq. 3.50 is an unsymmetric Lyapunov equation, with some rearrangement, X and U can be found by Gauss elimination.

Using estimated reference states, the control law is

$$u_k = -Kx_k + (U + KX)\hat{x}_{r,k} \tag{3.51}$$

Linear Quadratic Synthesis for a Time Varying Output

Consider the controller structure of Eq. 3.47. Another design approach would be to form a cost function, and to choose feedback gains K and feedforward gains F to minimize that cost.

We wish to minimize the cost function

$$J = \int_0^\infty (y - r)'Q(y - r) + u'Rudt \tag{3.52}$$

subject to the constraints of the dynamics of r and y (Eq. 3.46 and Eq. 3.26). Q and R are weights on the costs for the error and the control signal (Bryson, 1975; Gardner, 1984). [1] [5]

We first create an augmented state by putting x and x_r in parallel; defining

$$\vartheta_k \equiv [x'_k \; x'_{r,k}]'$$

and

$$\begin{aligned} \vartheta_{k+1} &= \begin{bmatrix} A & 0 \\ 0 & A_r \end{bmatrix} \vartheta_k + \begin{bmatrix} B \\ 0 \end{bmatrix} u_k \\ \vartheta_{k+1} &\equiv A_a \vartheta_k + B_a u_k \end{aligned} \tag{3.53}$$

By combining Eq. 3.26 and Eq. 3.46 into the cost function in Eq. 3.52, we form an expanded performance index

$$J = \int_0^\infty (\vartheta' Q_a \vartheta + u' R u) dt \tag{3.54}$$

constrained by Eq. 3.53. The expanded weight Q_a is

$$Q_a = \begin{bmatrix} C'QC & -C'QC_r \\ -C_r'QC & C_r'QC_r \end{bmatrix} \tag{3.55}$$

The minimization of this function results in an expanded control law

$$\begin{aligned} u_k &= -K_a \vartheta_k \\ &= -K_a [\, x_k' \; x_{r,k}' \,]' \end{aligned} \tag{3.56}$$

We can then separate the expanded gain K_a into a feedback gain K and a feedforward gain F. An estimator may be used for the reference input states (Eq. 3.48), giving a controller

$$u_k = -K x_k + F \hat{x}_{r,k} \tag{3.57}$$

Results and conclusions

Figure 3.6 shows the performance of all the controller designs. The performance of the design in section 4.3 ($\cdots$) is unacceptable. One might improve its performance by matching the parameters of the bandpass filter (Eq. 3.32) until the output matches the desired track, but the response of this feedforward augmentation scheme is essentially open loop.

The tracking performance of the other designs looks excellent, and the control efforts they demand are about the same. We therefore need to examine their responses when conditions are not perfect.

Figures 3.7 through 3.10 show the responses of the designs of sections 4.4 through 4.7 when (a) there is input and measurement noise, (b) the plant is not modelled exactly, (c) the input signal varies in amplitude and frequency, and (d) all these errors act at the same time. These errors are exaggerated to highlight the limitations of the controller. The noise simulated has a magnitude of 10% of the input magnitude; plant parameters are also varied by 10%, and the input amplitude and frequency are changed by 20%.

Figure 3.7 shows the responses of the design by plant inverse (section 4.4) with these errors present. This design doesn't really reject noise (Fig. 3.7a).

An error in modelling the plant shows how sensitive this design is (Fig. 3.7b). Larger errors in modelling, or drift in the plant parameters could easily cause the system to become unstable. Although this controller is easy to design, and especially easy to implement on a digital controller, its sensitivity makes it a poor candidate. This design is essentially an open loop design (although the gains are matched), and makes no use of the known characteristics of the reference input.

The controller design through "error space" (section 4.5, Fig. 3.8) shows good robustness with respect to variations in plant parameters (Fig. 3.8b) and to variations in the amplitude and frequency of the input (i.e. errors in modelling the input), as seen in Fig. 3.8c. This design's explicit use of the error signal, as a form of extended integral control, makes it somewhat sensitive to both input noise and measurement noise (Fig. 3.8a), but it performs very well otherwise.

The model following design presented in section 4.6 exhibits excellent noise rejection (Fig. 3.9a). When the input signal's frequency changes, (Fig. 3.9c) this design does not perform as well.

The behavior of the controller with feedforward gains designed by linear quadratic synthesis (section 4.7, Fig. 3.10) is very similar to the response with the model-following approach (Section 4.6). When the input is varied, this design has slightly better tracking performance than the model-following feedforward design, but it is slightly more sensitive to errors in the plant model and to noise. The similar behavior of these two designs is due to their similar structure: Compare Eq. 3.51 with Eq. 3.57.

The model-following feedforward and the linear quadratic designs exhibit superior noise rejection because they incorporate an estimator for the reference signal, which acts as a filter for the input signal. However, this adds complexity to the implementation, and at high sampling rates, the added computing burden for an additional estimator may cause problems.

It is instructive to look at the tracking error for these designs (Figure 3.11). The tracking performance (without noise) of the error space method is the best. Table 1 compares the performance of each design by its RMS tracking error. Despite its sensitivity to noise, the error space method (section 4.5) is still best. Although it will not track perfectly a varying input, its performance when there is modelling error and when the input signal changes (as would happen during machining, with tool forces acting on the spindle motor) is superior.

Changing the sampling basis from equal time intervals to equal spindle angle increments would eliminate the question of fluctuating input, at the cost of variability in the plant parameters.

Feedforward augmentation can be useful to extend the bandwidth of a system. While the design using the inverse of the plant can perform well, it is very sensitive. These two designs are best used only when nothing is known about the character of the input. The designs that use a model of the dynamics of the input are more robust, and can be extended easily to higher order inputs by just extending the model of the reference signal.

Table 3.1: RMS Tracking Errors

Method	plant error	varying input	plant + input var.	noise	all errors
Inverse	9.5813	0	3.6144	0.7882	3.8130
Error space	0.4797	2.3101	2.2785	1.2022	2.5450
Model follow	0.5233	2.8311	2.8373	0.5278	2.8503
Direct LQR	0.6911	2.6105	2.8001	0.7753	2.8499

The experimental results may change these conclusions, as they include effects not included in the simulations.

Bibliography

[1] Arthur E. Bryson and Yu-Chi Ho. *Applied Optimal Control.* Hemisphere, 1975.

[2] C. J. Chen. *A Laminar Flow Motor-Driven Machine Tool Spindle.* PhD thesis, Stanford University, 1985.

[3] P. J. Falter and T. A. Dow. Design and performance of a small-scale diamond turning machine. *Precision Engineering*, 9(4):185–190, 1987.

[4] Gene F. Franklin and Abbas Emami-Naeini. A new formulation of the multivariable robust servomechanism. Technical report, Stanford University Information Systems Laboratory, 1982.

[5] Bruce E. Gardner. *Feedforward/Feedback Control Logic for Robust Target Tracking.* PhD thesis, Stanford University, 1984.

[6] Ray H. Hashemi and Alan J. Laub. On the suboptimality of a parallel kalman filter. *IEEE Trans. on Aut. Contr.*, AC-33(2):214–217, 1988.

[7] Toshiro Higuchi, Takeshi Mizuno, Hiroshi Sugai, and Changjo Yun. Primary study on application of electro-hydraulic servo mechanism to noncircular cutting by a lathe. *Seisan Kenkyu (Tokyo Daigaku Seisan Gijutsu Kenkyujo)*, 36(2):71–73, February 1984.

[8] D. E. Luttrell and T. A. Dow. Development of a high speed system to control dynamic behaviour of mechanical structures. *Precision Engineering*, 9(4):191–200, 1987.

[9] S. R. Patterson and E. B. Magrab. Design and testing of a fast tool servo for diamond turning. *Precision Engineering*, 7(3):123–128, 1985.

[10] Raymond J. Roark and Warren C. Young. *Formulas for Stress and Strain.* McGraw-Hill, fifth edition, 1975.

[11] Michael David Sidman. *Adaptive Control of a Flexible Structure.* PhD thesis, Stanford University, 1986.

[12] S. P. Timoshenko and J. N. Goodier. *Theory of Elasticity*. McGraw-Hill, 1970.

[13] M. Tomizuka. Zero phase error tracking algorithm for digital control. In Max Donath, editor, *Dynamic Systems: Modelling and Control*. ASME, 1985.

[14] M. Tomizuka, M. S. Chen, S. Renn, and T. C. Tsao. Tool positioning for noncircular cutting with lathe. In *Proc. of the Amer. Contr. Conf.*, June 1986.

[15] Hy D. Tran and Daniel B. DeBra. Feedforward methods for non-circular turning on a lathe. In *Proc. of the 11th World Congress*. IFAC, 1990 (to be presented).

[16] Thomas L. Trankle and Arthur E. Bryson. Control logic to track outputs of a command generator. *J. of Guidance and Control*, 1(2):130–135, 1978.

[17] Taco J. Viersma. *Analysis, Synthesis, and Design of Hydraulic Servo Systems and Pipelines*. Elsevier Scientific Publishing, 1980.

[18] E. B. Wylie and V. L. Streeter. *Fluid Transients*. FEB Press, 1983.

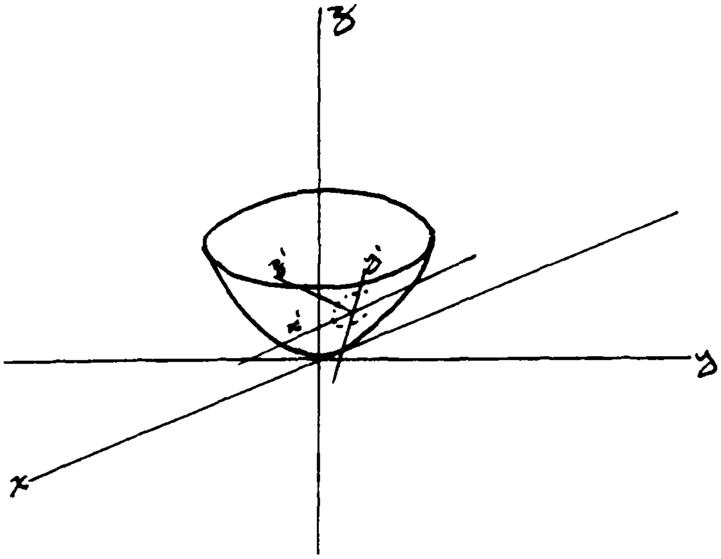


Figure 3.1: Off-axis paraboloid

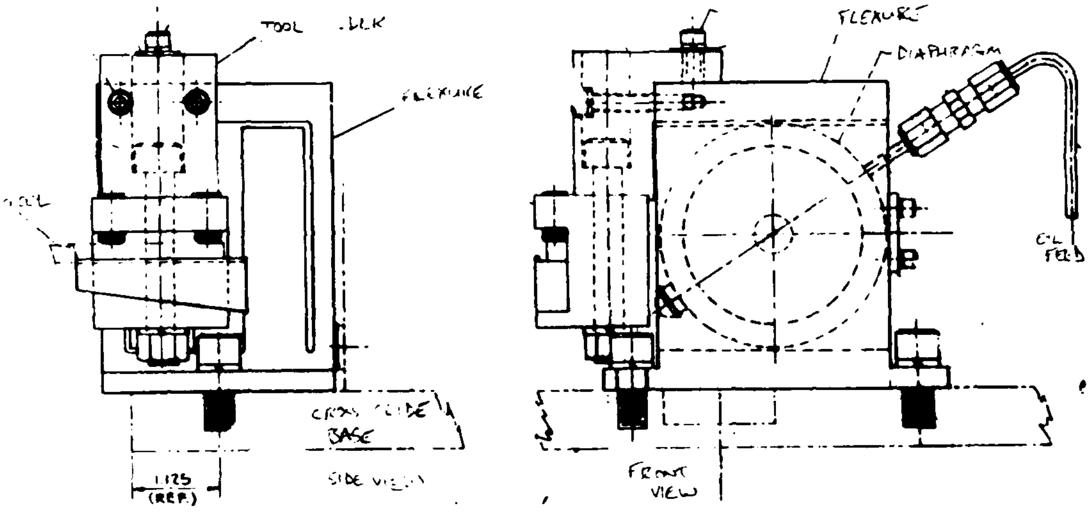


Figure 3.2: Actuator construction

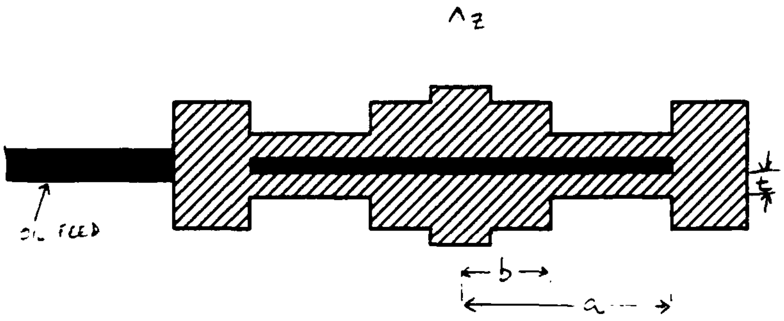


Figure 3.3: Diaphragm of short stroke actuator (cross-section)

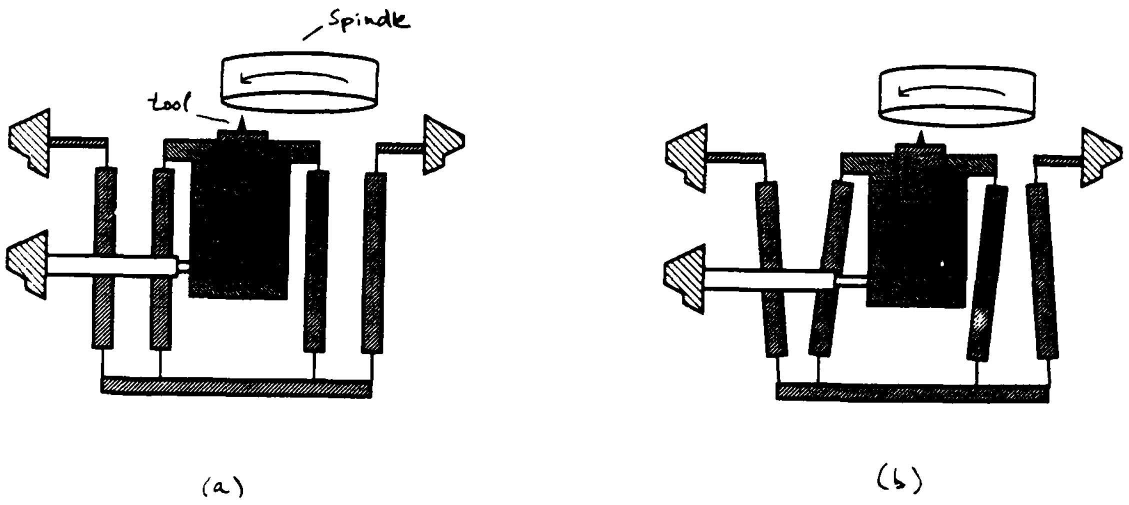


Figure 3.4: Flexure base for cross slide. (a) No translation. (b) Maximum translation.

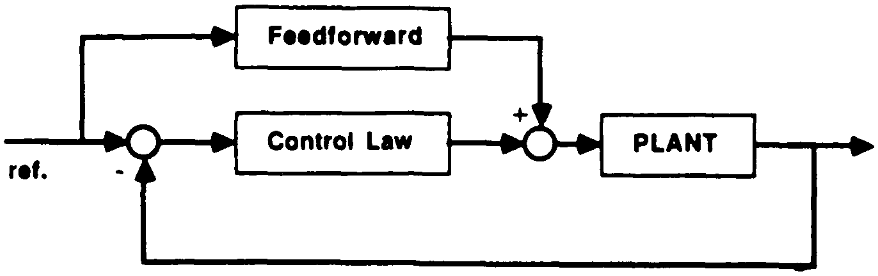


Figure 3.5: Using feedforward to speed transient response

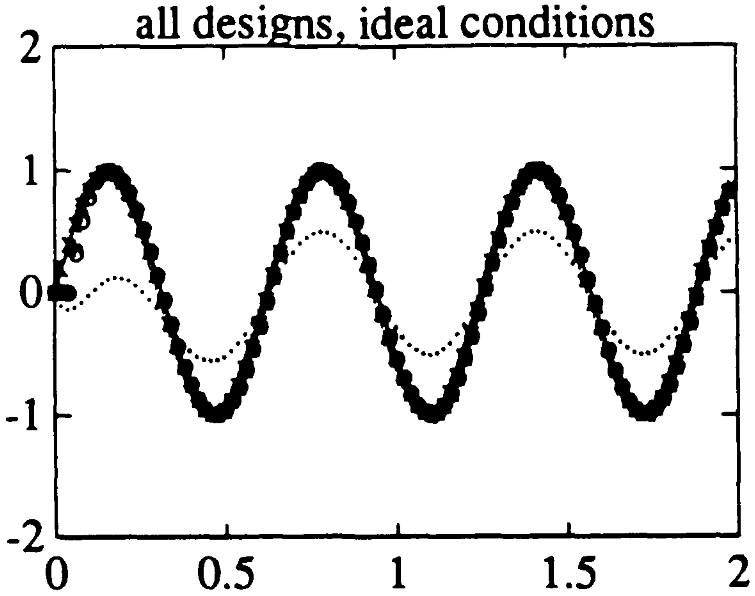


Figure 3.6: Response of all the designs to sin $10t$. — desire output; ··· feedforward augmentation; xxx inverse; tt +++ error space; ooo model following; *** LQR design

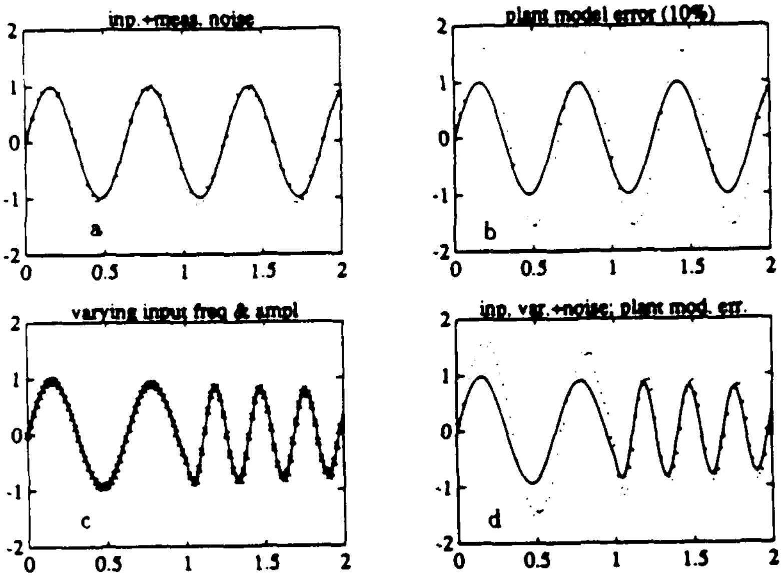


Figure 3.7: Response of the design via inverse (section 4.4). (a) meas. and inp. noise; (b) plant model error; (c) fluctuating input; (d) combined errors

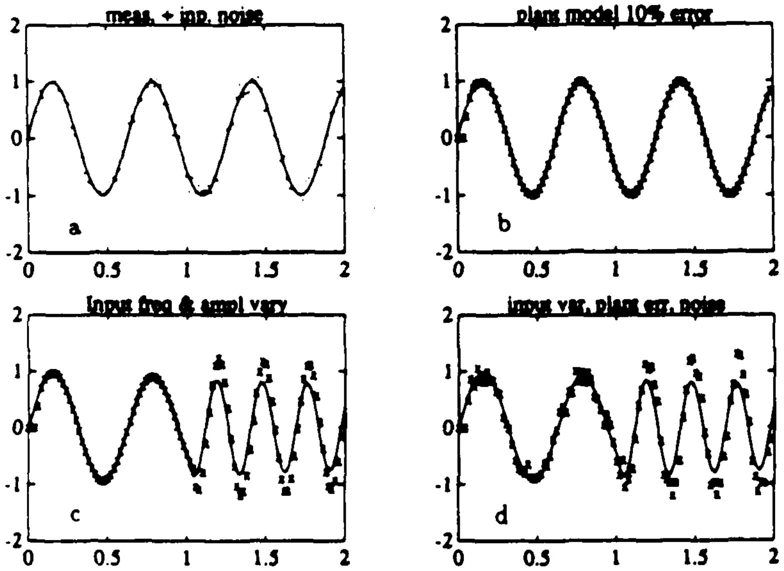


Figure 3.8: Response of the error space design (section 4.5). (a) meas. and inp. noise; (b) plant model error; (c) fluctuating input; (d) combined errors

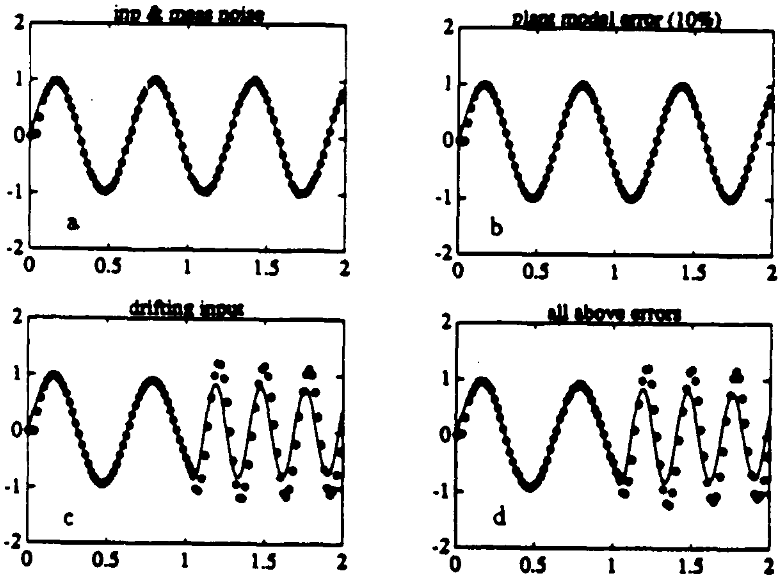


Figure 3.9: Response of the model following design (section 4.6). (a) meas. and inp. noise; (b) plant model error; (c) fluctuating input; (d) combined errors

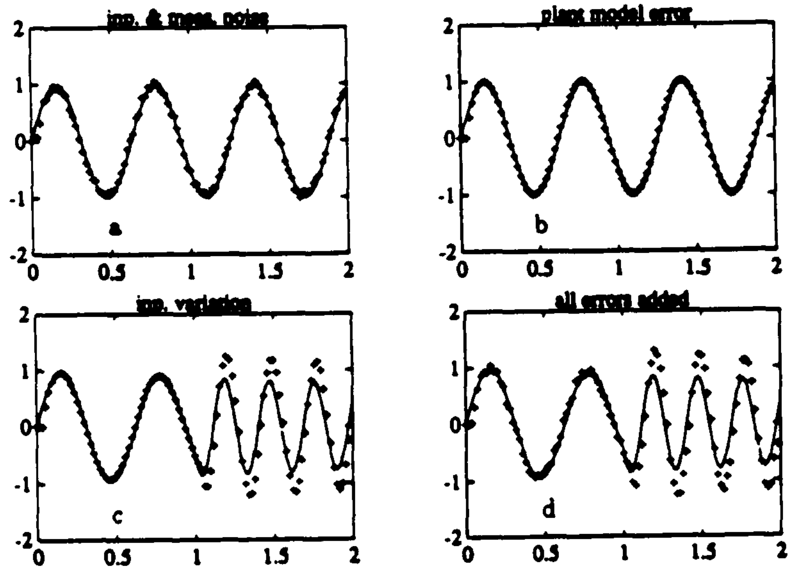


Figure 3.10: Response of LQR feedforward design (section 4.7). (a) meas. and inp. noise; (b) plant model error; (c) fluctuating input; (d) combined errors

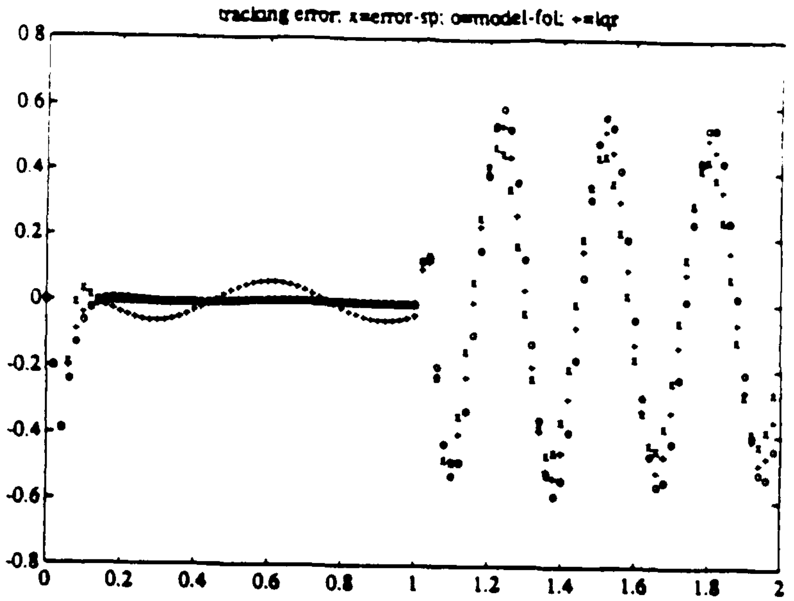


Figure 3.11: Tracking error of the various feedforward designs. Perfect model of plant, no noise. Input frequency is ramped, and amplitude has a step change at $t = 1$. x = error space design; o = model following design; + = LQR design.

3.5 LIQUID TEMPERATURE CONTROL

We have completed a theoretical and experimental development of temperature control employing a conventional heat exchanger. Feed forward of inlet coolant and controlled liquid temperature and feedback of the controlled liquid temperatures are studied to find the limits of complexity warranted for improves performance. The internal fluctuations of the heat transfer coefficient are found to be a fundamental limit in conventional heat exchangers.

3.5.1 Introduction

The need for high precision temperature control is well established. One millidegree celsius temperature changes in the structure of a machine tool with dimensions of the order of $1m$ in which the material has a coefficient of expansion of say $10^{-}5$ strain/degree celsius causes a change in dimension of $10nm$. This is as much as one is willing to allot to thermal effects in a machine designed to have accuracy of the order of $35nm$ - a representative number for example for diamond turning machines [Bryan, 1979]. The development of temperature controllers for machine tools has evolved with the machines in a systematic way with may contibutions [Kraakman, deGast], [Gijsbers], [Bryan, et al, 1982], [Brown, et al] which draw on a broader application of heat exchangers, e.g. [Masubuchi], [Giles], [Gartner, Harrison], [Gartner, Daane], [Holman], [Sieder, Tate], [Baur, Isermann], [Kays, London, 1984]. We have addressed the question of controolling the temperature of oil used in a precision machine with chill water.

The performance of fluid temperature controllers has been accomplished well at low frequencies using integral control; however, the performance of the integral controller degrades as the fequency of the disturbance signal increases. There is a need to increase the bandwidth of control, so that the effect of medium frequency disturbances can be reduced.

The major contribution of this research is the extension of the frequency range of the precise fluid temperature control though a combination of feedforward and feedback control. The full results are given in [Chou, 1988] and in this form will be presented this summer at the CIRP [Chou et al, 1990]. The contributions are in three areas: the identification of a heat exchanger, the design of the control laws, and suggested methods of eliminating the oil temperature fluctuation above the control system bandwidth.

3.5.2 System Description

The schematic diagram of an oil temperature contol system is shown in Fig. 1. Starting at the machine sump, the oil flows through a filter located on the suction side of the pump, through the pump to a heat echanger (Young Radiation F-303-EY-4P-B), onto the machine, and back to the sump. At the same time, the chilled water supplied by a chiller flows through a servovalve (Hydraulic Servo Control Model 70) to the heat exchanger, and back to the chiller where the water is cooled again by refigeration. Installed at the inlets and outlets of the oil and water flow passages to the heat exchanger, and back to the chiller where the water is cooled again by refigeration. Installed at the inlets thermistors, which are used to sense the oil and water temperatures. The temperature signals are communicated to the computer through interface circuits, which consist of conditioners and analog-to-digital converters. The servovalve is used to regualte the water flow rate under the command signal issued from the computer, through a digital-to-analog converter and the power amplifier (the driver).

3.5.3 Identification of a Heat Exchanger

An experimentally determined system model representing the dynamic response of a commercially available heat exhanger was established. This system consists of two disturbance models with inlet oil and water temperature variations as the disturbance inputs, and one plant model with the water flow rate as the control input. The output of the system is the outlet oil temperature variation.

Heaters were used to vary the inlet oil and wate. temperaures so that at least the low frequency modes of the distruvance models were excited. An iterative identification scheme was conceived to estimate the parameters in each of the disturbance models.

The plant model was identified by using the relationship between the plant input and the plant output which was obtained by subtracting the disturbance outputs from the system output.

The heat exchanger contains 76 tubes, each having a diameter of 1/4 inches (6.35mm) and a length of 18 inches (0.457m). The tubes are divided into four passes, so each pass has 10 tubes in parallel. The total heat transfer area is $0.69m^2$. The shell volume is divided into four compartments by three baffles, which direct the oil flow so as to obtain a maximum heat transfer with minimum fluid pressure drop. The shell side and the tube side volume are approximately 1.0 gallons and 0.5 gallons, respectively. At the nominal 40 gal/min (2.5 1/sec) oil flow rate and 1 gal/min (63 mlsec) water flow rate, the dwell times are 1.5 seconds and 30 seconds for oil and water, respectively. The nominal operating temperatures are $68 \deg F$ $(20 \deg C)$ for oil and 50 $\deg F$ $(10 \deg C)$for water.

System Identification

An analytical approach for modelling a heat exchanger is limited. Where solutions exist they suggest a complicated combination of time delay, lag and distributed properties [Masubuchi], [Gilles], [Gartner, Harrison], [Gartner, Daane]. We confirmed this with a single tube model. The Young heat exchanger we used has 76 tubes and was too complicated to model analytically so we chose to fit a model to experimental data. The output contains a wide band noise of significant amplitude so we chose a parametric fit using least squares fitting with the data prefiltered [Wahlbert, Ljung, 1986], [Ljung, 1987], Sidman], [Rovner]. The model is shown in Fig. 3.5.2. Two of the transfer functions represent reponse to disturbances and are used for feed forward control. The third is the reponse to control inputs.

Candidates for System Models

After the system models have been established, we proceed to the second step of system identification, the assignment of candidates for each ofthe individual models. Although we found analytically that the disturbance model $G1(s)$ of a single-tube heat exchanger can be approximated by constant gain with a time delay, and $G2(s)$ can be approximated by a first order system perhaps with a time delay, and $G2(s)$ can be approximated by a first order system perhaps with a time delay, we are not quite sure whether or not the system model of this complex heat exhanger can be approximated the same way. To allow for more complexity, we assigned several rational transfer functions of different orders for each of the individual models expressed in polynomials of s or z. After a careful examination of some preliminary data, we decided that the plant model should be approximated by a first, second, or third order system with a time delay, and the two disturbance models should be approximated by a zero, first, or second order system with a time delay. Although a higher order system is possible, usually a high order system can be approximated by a lower order system with a time delay. It is important to note that the exact model of this complex heat exhanger is not known, and will certainly not be a function of transcendental functions, as we found in our analytical model of a single tube heat exchanger, or it would be an even more complicated function [Chou]. Therefore, the higher order model is not necessary if a lower order model can successfully describe the dynamic response of the real system.

The time delays of the two disturbance models can be justified from the facts that it takes time for the oil flow from the inlet to the outlet, and for the water to travel from one end of the heat exchanger to the other end. The delay time will be proportional or equal to the transport time.

Summarizing the above descriptions, the possible candidates forthe system model are presented in Table 3.5.1. Here the parameters b*12*, b*22*, and b*32* included in the zero and first order models are necessary when the time delay is not an integer number of the sampling period. For details on transforming a continuous model with a time delay into a discrete model, see Franklin and Powell [1980].

Test Inputs

The signal-to-noise ratio of the outlet oil temperature plays an important role in the accuracy of the identified parameters. A low signal-to-noise ratio signal will lead to inaccurate results. The typical temperature noise level at the oil outlet is 0.01deg F (5.6 m deg C). We need a signal-to-noise ratio greater than ten. Taking this consideration into account, the amplitudes of the input temperature excitation applied at both oil and water inlets were chosen to be 0.16 deg F(0.089 deg C) and 4.2 deg F(2.3 deg DC), respectively, which resulted in 0.14 deg F (0.078 deg C) and 0.12 deg F (0.067 deg C) temperatures variations at the oil outlet. The signal-to-noise ratios were roughly 14 1n3 12, which are larger than ten and are believed to be big enough to distinguish the signal from the noise. Note that we do not want the inlet oil temperature variation to be too large because the oil may burn at the heater surface if a high heat rate is applied. The typical temperature variations of the inlet oil, inlet water, and outlet oil are shown in Fig. 3.5.3. Here we can only clearly see the one-to-one correspondence of temperature variations between the inputs and the output. The first and third outlet oil temperature variations are caused by the two inlet oil temperature pulses, and the second and fourth outlet oil temperature variations are caused by the two inlet water temperature pulses.

Test input models of zeroth, first and second order with time delays were tested by comparing the experimental outputs with simulated results using the model. It was found that zeroth order for the oil disturbance and first order for the water disturbance were adequate. Higher order models did not improve significantly the matching of the experimental and simulated results. See Fig. 3.5.4. The proportionality of these functions is a weak function of the water flow. Over the full range there is approximately a ±15% for the water disturbance transfer functions.

The control response was modelled using a step input. A first order model with time delay was found to be adequate. See Fig. 3.5.5 for the model fitted without prefiltering and Fig. 3.5.6 with prefiltering.

The models are summarized in Table 3.5.2.

3.5.4 Control

We have found that classical proportional plus integral (PI) plus feedforward (FF) was slightly more successful than a linear quadratic regulator with integral control and feed forward (LQR). The PI is shown in Fig.3.5.7 and the LQR in Fig.3.5.8. (After the high frequency mixing discusses in the next paragraph.)

While in theory the PI controller with feedforward compensation is able to control the output of the system perfectly even in the presence of disturbances, the performance of this control law is limited by the outlet oil temperature fluctuations.

The oil temperature fluctuations are probably the results of unsteady flow that occurs when fluid flows across a tube, the incomplete mixing of the fluid thermal-boundary that forms after heat has been added or removed from the fluid, and the secondary and leakage flows of the oil through the baffles inside the heat exchanger. To illustrate this quantitatively, assume the temperature difference between the oil (20 $\deg C$) and the chill water (10 $\deg C$) is 10 $\deg C$, then a change of only 10^{-3} in the heat transfer coefficient can cause a change in the oil outlet temperature of $10m \deg C$ even if the oil and water inlet temperatures remain constant. So it is not surprising that the observed oil outlet temperature fluctutations exist. A typical frequency spectrum for the outlet oil temperature is pictured in Fig. 3.5.9, which indicates that most of the temperature fluctuations have frequency higher than the system handwidth whose value is determined by the plant time constant and time delay and is roughly equal to $2 \bullet 10^{-2} Hz$.

Since the temperature fluctuations wtih frequency higher than the system bandwidth cannot be improved by feedback, some methods or devices ought to be found to reduce the fluctutations. A simple way to achieve this goal is to use elbows, which generate secondary flow and vortices around corners, and yield a better mixing of the fluid. London [1988] indicates in his experience that several consecutive elbows arranged in different orientations (planes) have better performance than the same number of elbows arranged in different orientations (planes) have better performance than the same number of elbows arranged in only one orientation. Our outlet oil temperature sensor was originally installed right at the exit of the heat exchanger before the elbows, and therefore sensed strong temperature fluctuations. After realizing the effect of the elbows, we installed another thermistor some distance downstream of three existing elbows that were arranged in two different planes. The measured outlet oil temperature at the new location indicates that a temperature fluctuation of $2m$ deg F$(1.1m \deg C)$ is achieved (this can be estimated from Figs. 3.5.10 and 3.5.7), which is about an order of magnitude smaller than that without downstream mixing. The typical frequency spectrum of the outlet oil temperature with fluctuations reduced is illustrated in Fig. 3.5.11, showing a decrease in the spectrum magnitude in the high frequency region when compared with Fig. 3.5.9.

An on-off controller was implemented for comparison. Its average value could be made comparable to the PI controller but a typical limit cycle reduces the short term performance (see Fig. 3.5.12). A delay line mixer was investigated but not tested. If half of the flow is delayed so that its minimum temperature arrives when the other path is at maximum temperature, one can average out the limit cycle. However, as disturbances change the duty cycle and the delay is not correctly matched, the minimal conditions don not match and the effectiveness of the delay line compensator reduces (see Fig. 3.5.13). A four delay line system is more robust but more complicated.

3.5.5 Conclusions

For the heat exchanger used, the disturbance models should be identified before the plant model, because the control input can be held fixed so that it will not affect the output, leaving the output influenced only by the two disturbance inputs. Because of the non-separable property of the disturbance outputs, a numerical method called iterative least square identification is effective to identify the two disturbance models simultaneously.

The slight nonlinerarity of the plant dynamics can be neglected if only stability is concerned. It causes a small error in the output, if a fixed plant model is used to implement the feedforward compensation alogrithm which is used to eliminate the disturbance effect. This nonlinearity can be fitted very well by a third order polynomial.

The response time of the control system, including the characteristic time and delay time, limits the bandwidth of feedback control. In our heat exchanger with a response time of 10 seconds feedback can deal with output errors up to about 0.1rad/sec. Integral feedback is very effective in the very low frequency region; however, its performance degrades as the frequency content of the distrubance increases. Feedforward compensation is effective in reducing the effect of the inlet oil and water temperature disturbances with frequency up to the bandwidth of the physical system, whereas the control bandwidth is limited by the time delay. However, the effectiveness of the feedforward compensation is inferior to the integral feed back in the extremely low frequency region if there are considerable modeling errors. The combination of integral feedback and feedforward compensation can deal with the entire frequency region up to the physical system bandwidth. Above this frequency residual input disturbances and compensated output noise must be smoothed by some type of averaging or by using a different type of heat exchanger.

For the commercially available high effectiveness heat exchanger, oil temperature fluctuation is unavoidable; however, we can exploit some methods to reduce the temperature fluctuation. Two kinds of approaches are proposed. The first approach is to introauce devices that can increase the fluid mixing. These devices include elbows, baffles, and equalizers. The second approach is to use a liquid-coupled indirect-transfer-type heat exchanger to reduce the temperature gradient in the thermal boundary.

REFERENCES

Bauer, U., Isermann, R., 1980, "One-Line Identification of a Heat Exchanger with a Process Computer- A Case Study," *Automatica*, 16:487-496.

Bryan, J.B., 1979, "Design and Construction of an Ultra Precision 84-inch Diamond Turning Machine", *Precision Eng.* 1:1.

Bryan, J.B., et al, 1972, "A Practical Solution to the Thermal Stability Problem in Machine Tools", S.M.E. Tech. Paper, Dearborn, MI USA, No. MR72-138 .

Chou, C., (Thesis), Stanford University, Stanford, CA 94305

Chou, C., and DeBra, D.B., 1990, "Liquid Temperature Control for Precision Tools", to be presented at CIRP in Berlin, FRG, August.

DeBra, D.B., Victor, R.A., Bryan, J.B., 1986, "Shower and High Pressure Oil Temperature Control", *CIRP Annals*, 35:1.

Franklin, G.F., Powell, J.D., (1980), "Digital Control of Dynamic Systems", Addison-Wesley, MA, USA.

Gartner, J.R., Harrison, H.L., 1965, "Dynamic characteristics of Water-to-Air Crossflow Heat Exchangers"'*ASHRAE Trans.*, 71:I:212-214.

Gartner, J.R., Daane, L.E., 1969, "Dynamic Response Relations for a Surpentine Crossflow Heat Exchanger with Water Velocity Disturbance," *ASHRAE Trans.*, 75:I53-68.

Gijsbers, T.G., 1980, "COLATH, a Numerical Controlled Lathe for a Very High Precision",*Philips Tech. Rev.*, 39:9:229-245.

Gilles, G., Sept. 1974, "New Results in Modeling Heat Exchanger Dynamics",*J. of Dynamic Syst., Meas., and Control, Trans. of ASME*, 96:3"G:277-282.

Holman, J.P., 1981, *Heat Transfer*, McGraw-Hill, New York, NY, USA.

Kays, W.M., London, A.L., 1984, *Compact Heat Exchanger*, McGraw-Hill, New York.

Kraakman, H.J.J., deGast, J.G.C., 1969, "A Precision Lathe with Hydrostatic Bearing and Drive",*Philips Tech. Rev.*, 30:5:117-133.

Ljung, L., 1986, *System Identification-Theory for the User, Prentice-Hall,* *Inc. New Jersey, USA.*

London, A.L., 1988, Private Communication.

Masubuchi, M., Mar., 1960, "Dynamic Response and Control of Multipass Heat Exchangers", J. Basic Eng., Trans. of the ASME, *82:D:51-65.*

Roblee, J.W., 1985, "Precision Temperature Control for Optical Manufacturing", Second Intntl. Tech. Symp. on Optical and Electro-Optical Appl. Sci. and Eng.

Rovner, D.M., 1987, *Experiments in Adaptive Controlof a Very Flexible One Link Manipulator*, Ph.D. Thesis, Stanford University, Stanford, CA, USA.

Sidman, M.D., 1986, *Adaptive Control of a Flexible Structure*, Ph.D. Thesis, Stanford University, Stanford, CA, USA.

Sieder, E.N., Tate, C.E., 1936, "Heat Transfer and Pressure Drop of Liquids in Tubes", *Ind. Eng. Chem.*, 28:1429.

Wahlberg, B., Ljung, L., 1986, "Design Variables for Bias Distribution in Transfer Function Estimation", *Trans. on Automatic Cont.*, AC-31:2:134-144

FIGURES

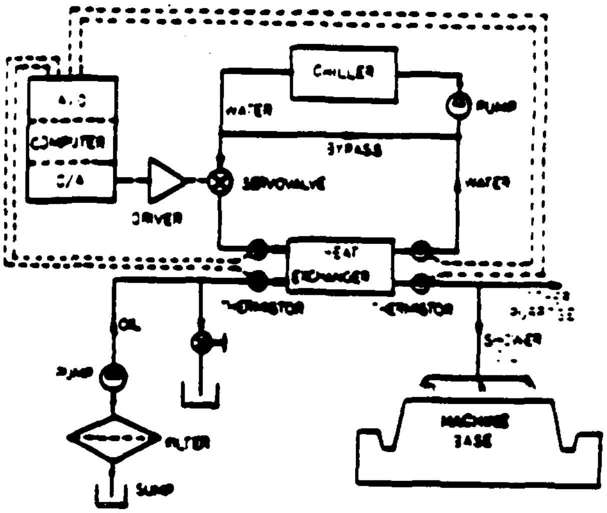


Fig. 3.5.1 **Schematic Diagram of the Oil Shower Temperature Control for a Precision Machine.**

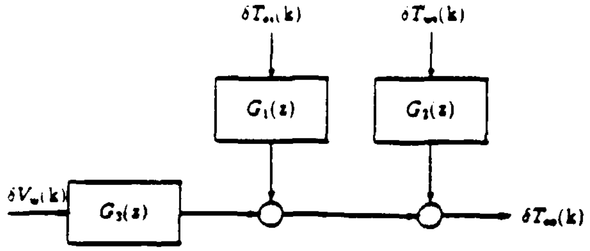


Fig. 3.5.2 Model Structure for the Heat Exchanger. $\delta T_{oi}(k)$ is the inlet oil temperature variation, $\delta T_{wi}(k)$ is the inlet water temperture variation, $\delta T_{oo}(k)$ is the outlet oil temperature variation, δV_w is the water flow velocity variation.

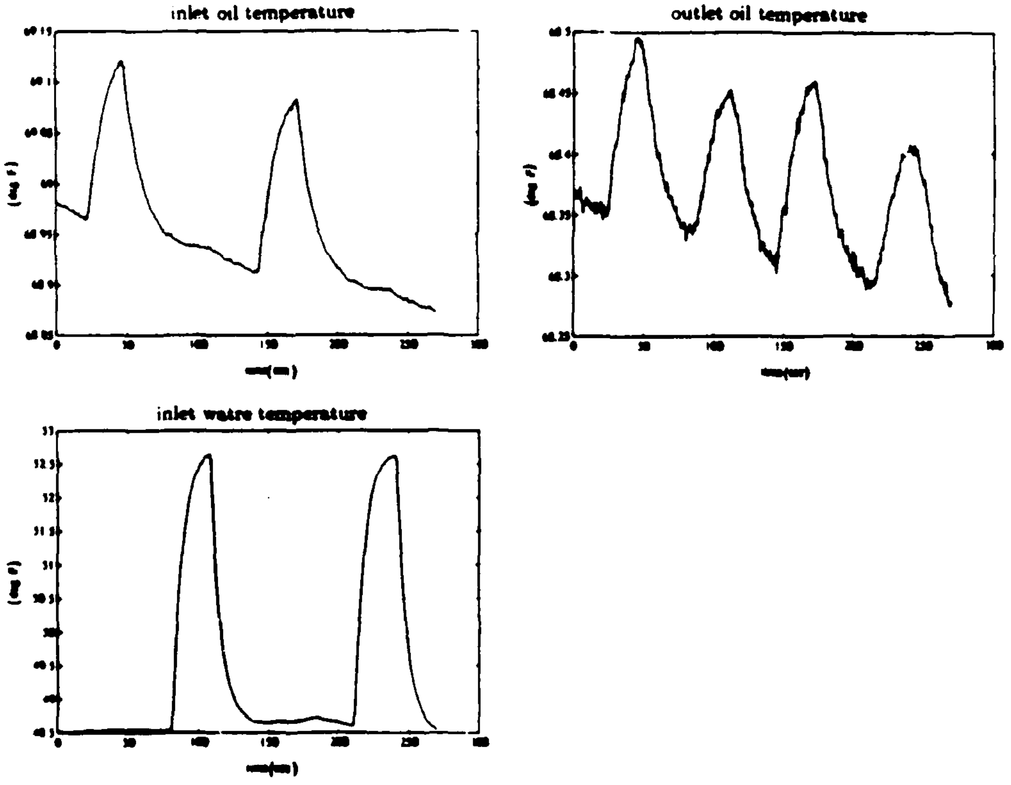
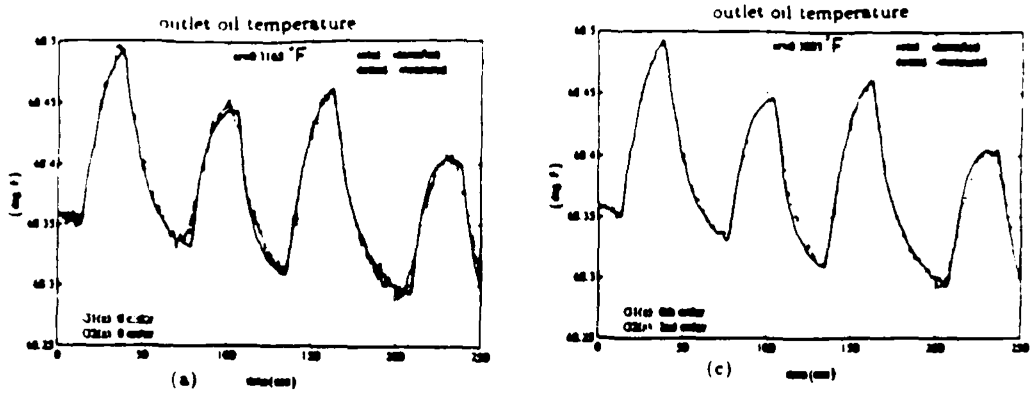


Fig. 3.5.3 Typical Temperature Variation of Inlet Oil, Inlet Water, and Outlet Oil for Disturbance Model Identification.



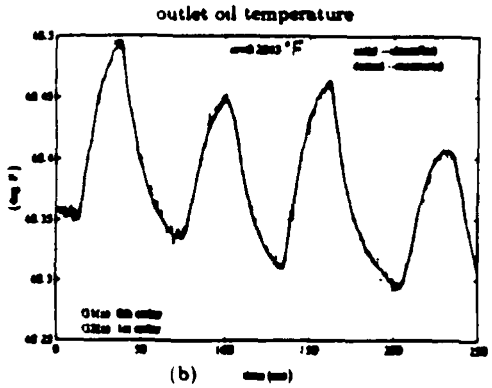


Fig. 3.5.4 **Measured and Simulated Outlet Oil Temperature.**

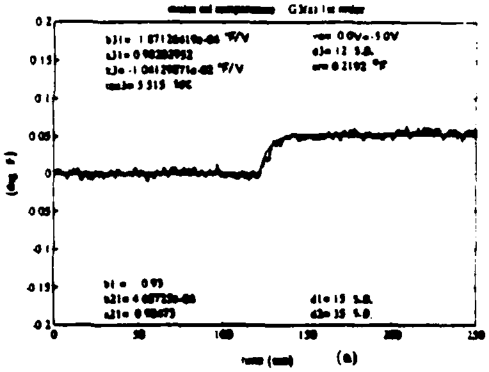


Fig. 3.5.5 **Measured and Simulated Plant Output.**

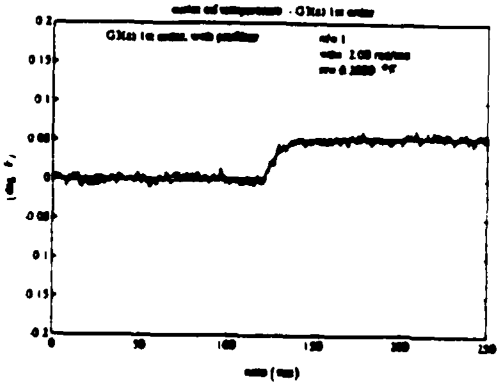


Fig. 3.5.6 **Measured and Simulated Plant Output y_3. Prefilter Applied.**

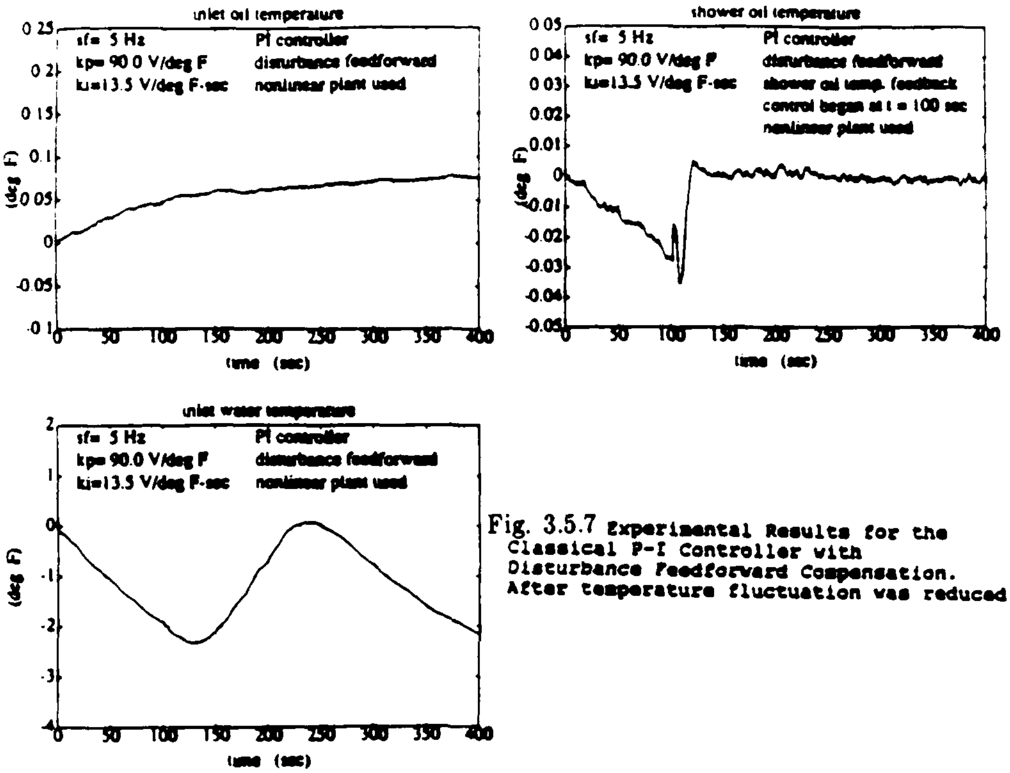


Fig. 3.5.7 Experimental Results for the Classical P-I Controller with Disturbance Feedforward Compensation. After temperature fluctuation was reduced.

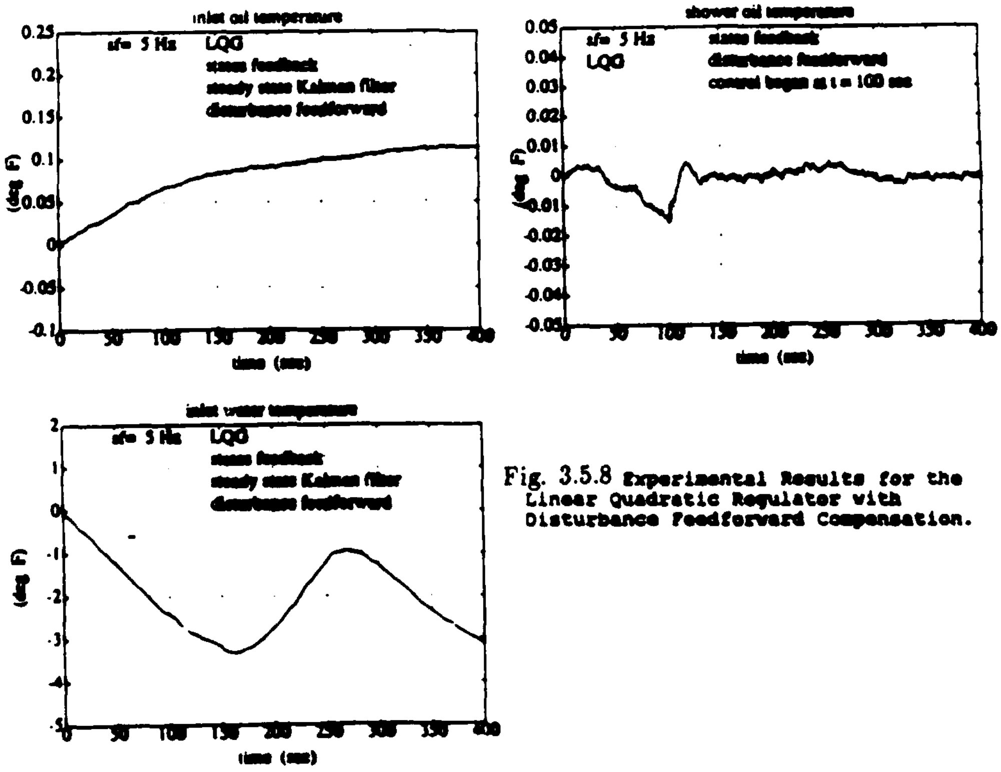


Fig. 3.5.8 Experimental Results for the Linear Quadratic Regulator with Disturbance Feedforward Compensation.

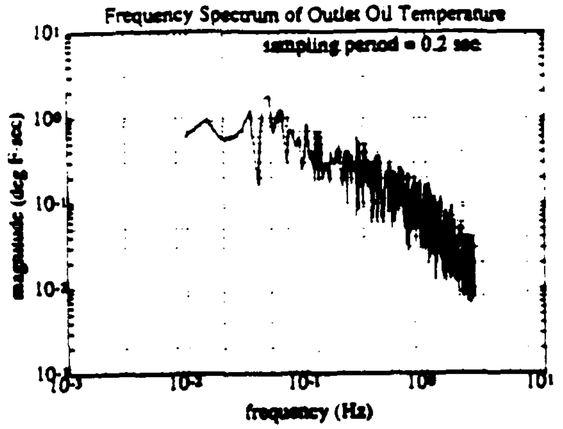


Fig. 3.5.9 Typical Frequency Spectrum of the Outlet Oil Temperature. Before temperature fluctuation was reduced.

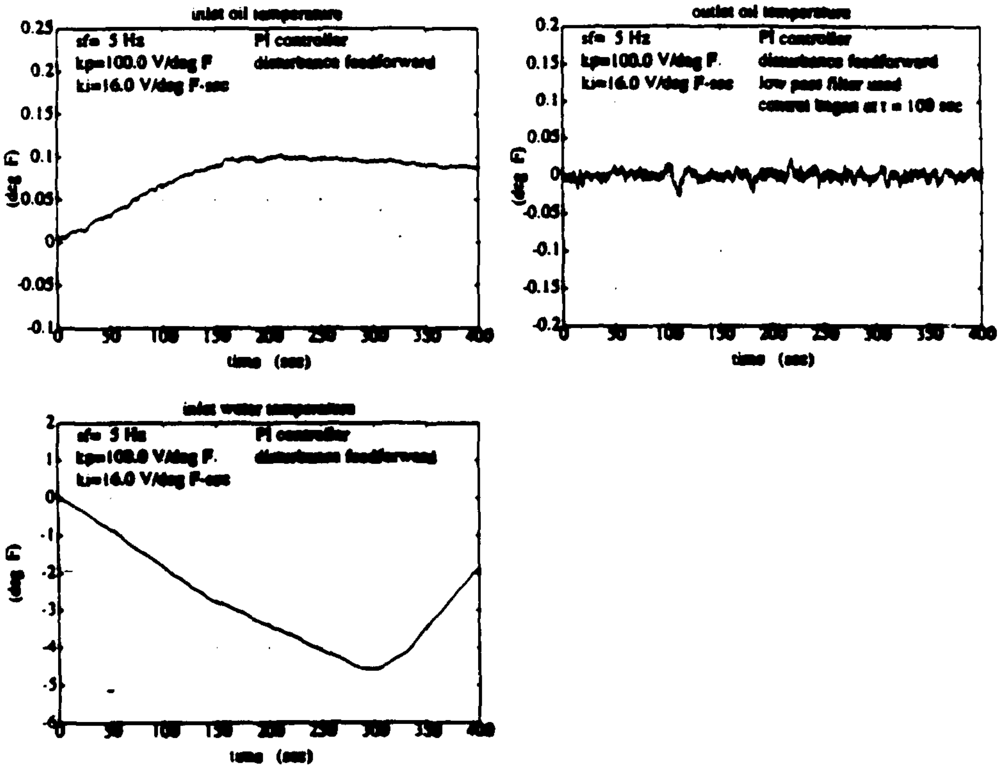


Fig. 3.5.10 Experimental Results for the Classical P-I Controller with Disturbance Feedforward Compensation. Before noise was improved.

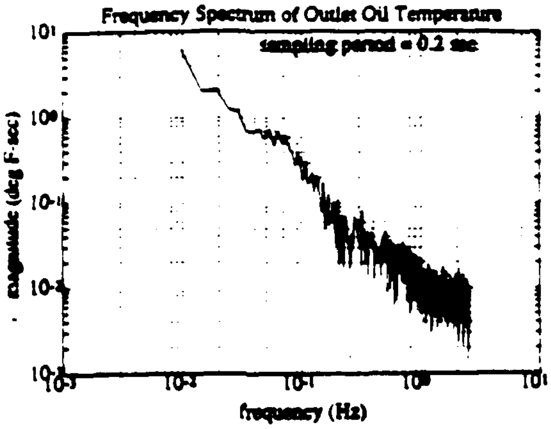


Fig. 3.5.11 Typical Frequency Spectrum of the Outlet Oil Temperature. After temperature fluctuation was reduced.

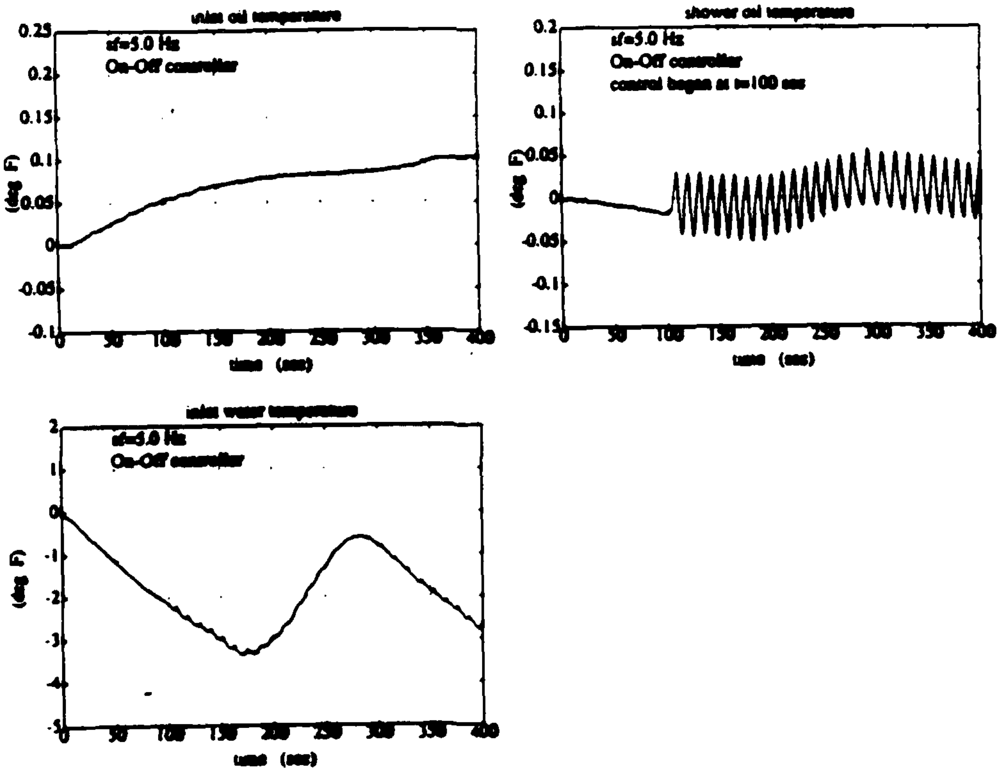


Fig. 3.5.12 Experimental Results for the On-Off Controller with Zero Hysteresis.

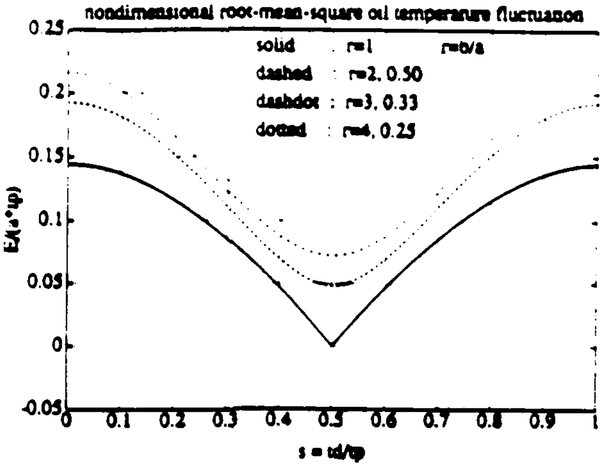


Fig. 3.5.13 Nondimensional Root-Mean-Square Residual Temperature Fluctuation for a Two-Line Delay Line Mixer.

	$G_1(z)$	$G_2(z)$	$G_3(z)$
Zero order	$(b_{11}z^{-1} + b_{12}z^{-2})z^{-d_1}$	$(b_{21}z^{-1} + b_{22}z^{-2})z^{-d_2}$	—
First order	$\frac{(b_{11}z^{-1} + b_{12}z^{-2})z^{-d_1}}{1 - a_{11}z^{-1}}$	$\frac{(b_{21}z^{-1} + b_{22}z^{-2})z^{-d_2}}{1 - a_{21}z^{-1}}$	$\frac{(b_{31}z^{-1} + b_{32}z^{-2})z^{-d_3}}{1 - a_{31}z^{-1}}$
Second order	$\frac{(b_{11}z^{-1} + b_{12}z^{-2})z^{-d_1}}{1 - a_{11}z^{-1} - a_{12}z^{-2}}$	$\frac{(b_{21}z^{-1} + b_{22}z^{-2})z^{-d_2}}{1 - a_{21}z^{-1} - a_{22}z^{-2}}$	$\frac{(b_{31}z^{-1} + b_{32}z^{-2})z^{-d_3}}{1 - a_{31}z^{-1} - a_{32}z^{-2}}$
Third order	—	—	$\frac{(b_{31}z^{-1} + b_{32}z^{-2} + b_{33}z^{-3})z^{-d_3}}{1 - a_{31}z^{-1} - a_{32}z^{-2} - a_{33}z^{-3}}$

Table 3.5.1: Candidate Models for the System Model of a Heat Exchanger

Models	Parameter Values						Reference
$G_1(s) = b_1 e^{-t_{d_1}s}$	b_{11}		d_1 (s.p.)	b_1		t_{d_1} (sec)	Table 3.2
$G_1(z) = b_{11}z^{-(d_1+1)}$	0.938		15	0.938		1.8	
$G_2(s) = \frac{b_2 e^{-t_{d_2}s}}{1+\tau_2 s}$	b_{21}	a_{21}	d_2 (s.p.)	b_2	τ_2 (sec)	t_{d_2} (sec)	Page 58 [Chou]
$G_2(z) = \frac{b_{21}z^{-(d_2+1)}}{1 - a_{21}z^{-1}}$	0.000483	0.988	27	0.0316	6.48	2.8	
$G_3(s) = \frac{b_3 e^{-t_{d_3}s}}{1+\tau_3 s}$	b_{31} (°F/V)	a_{31}	d_3 (s.p.)	b_3 (°F/V)	τ_3 (sec)	t_{d_3} (sec)	Table 3.3 Page 72
$G_3(z) = \frac{b_{31}z^{-(d_3+1)}}{1 - a_{31}z^{-1}}$	−0.0001876	0.983	11	−0.01153	5.86	1.2	[Chou]

Table 3.5.2: Summary of Nominal Models for the Heat Exchanger. u = 0 V, s.p. = sampling period = 0.1 sec. [from Chou 1988]

Chapter 4

COMPUTING ENVIRONMENT

4.1 MONITORING SYSTEM FOR A PRECISION TURNING MACHINE

Diamond turning machines are used to manufacture parts to extreme precision and surface finish quality. Consequently, the turning machine itself is more subtle than most conventional machine tools and requires sophisticated monitoring to indicate the quality and success of the machine's operation. This report describes work on a monitoring system for the Stanford Diamond Turning Machine.

Classical monitoring systems use a single filtering or signal processing algorithm to extract failure sensitive features from the data, most commonly FFT or Kalman filter. Monitoring strategies based on FFT spectrum estimates are often empirical and capable only of generating alarm without diagnostic information. Kalman filter based methods offer possibilities for diagnosing abnormalities but suffer from sensitivity to modeling errors. In our research, a monitoring strategy has been developed that is based on estimation of meaningful physical parameters, allowing integration of information from diverse sources with indication of the quality of the estimates. This enables the incorporation of practically any information into the estimate of a physical parameter, including prior expectations, multiple sensor inputs and historical statistics.

The underlying mathematics are based on the use of probability density functions as a representation of the available information. The rules of Bayesian analysis then allow the extraction of unknown parameters using a classical procedure known as "maximum a posteriori" estimation. An efficient algorithm has been developed to carry out the estimation. Examples taken from the diamond turning machine are given.

4.2 INTRODUCTION

Rotating machinery lends itself particularly well to monitoring because it is easily understood and relatively simple methods and commonly available instruments go a long way. Typically this would involve measurement of vibrations or shaft motions, sometimes followed by spectrum analysis and evaluation of certain features of the spectrum. Commonly available instruments help automating the process of acquiring signals and estimating their spectrum, followed by a comparison to a reference spectrum from a healthy machine. Almost all monitoring of rotating machinery is currently based on fast Fourier transform (FFT) spectrum estimates of signal spectra, most likely because of the simplicity of the procedure and the availability of FFT analyzers at reasonable prices. In academia and to a limited extent in industry, more exotic methods have been explored, some of which are described later.

Much attention has been devoted to the monitoring of machine tools where production efficiency is at stake and failures such as tool breakage are common. This is also where the flora of methods is the richest, with examples found in [4, 11, 35]. Other applications include fault detection and diagnosis in process industries [39, 21, 46], the monitoring of aircraft for increased safety [16, 47], increasing reliability and efficiency of turbomachinery [3] and monitoring of electrical motors [17, 26].

Monitoring systems that involve extensive analysis of measurement data have been made feasible by the common availability of computers and instrumentation at reasonable cost. They have usually been developed by end users of production machinery to boost efficiency and reliability. Consequently, many monitoring systems that are reported show the signs of being based on empirical observations which in many cases predate the computers. Having a computer carry out these observations continuously then constitutes a monitoring system in its simplest form. As noted earlier, this approach is exemplified by the extensive use of the fast Fourier transform applied to a single data stream with subsequent examination of power spectral density at predetermined frequencies. Interestingly, the rapid gains in computer technology have not led to corresponding advances in monitoring techniques — the basic pattern is the same, only with higher data rates, more tests per unit time and a more sophisticated user interface [30]. Efforts to develop the monitoring technology with the opportunities afforded by the increased processing capabilities have mostly been confined to academia and research oriented companies with little transfer to the factory floor. The most likely explanation is that the engineering staff of the maintenance department that usually are responsible for developing a monitoring system can hardly be expected to master all the different disciplines that must be incorporated into a more sophisticated system. A monitoring system akin to the one introduced here is more likely to be successfully developed in the design department of the machine's manufacturer rather than with the end user and should also be developed alongside the machine itself. The technical reasons for this will become clear later but the question of economic feasibility of this proposition is left open.

Since this research is partly intended to point in a feasible direction of future development of monitoring systems, it is proper to review some of the systems and methods reported so far.

4.3 REVIEW OF MONITORING TECHNIQUES

4.3.1 Methodologies

The monitoring systems reported in the literature can be classified into three categories by the underlying methodology. The principal difference between the three lies mostly in the way information on known system properties is introduced.

The simplest alternative is of course not to introduce any such information and make no assumptions about structure or causal relationships in the system. In that case, information is gathered from the system when it is in good working condition (or similar systems that work well) and that compared to later observations to reveal deviations from normal. A case in point is the practice of generating topographical maps of brain activity based on electroencephalograms. Brain activity of a group of healthy people, sometimes in response to certain sensory stimuli, is examined and compared to that of a patient under observation. Certain disorders, such as epilepsy, dyslexia and some sensory disorders can then be diagnosed based on the way the patient's responses deviate from the control grouף's. The reason for this approach is that the mechanisms which generate the electrical activity in the brain are not understood and hardly observable. This approach is also often used in monitoring of machinery, such as in the common case of comparing vibration spectra to older ones that are known to be characteristic of good condition.

The second methodology involves making assumptions about the structure of the system or the observed signals which are not based on analysis of the system but are instead suggested by the methods available for efficiently reducing the information to a small set of descriptive numbers. These numbers can only indirectly be related to physical properties of the system. This approach, which is often referred to as "black box modeling" is therefore often efficient in terms of computations and requires no prior analysis but interpretation of the resulting numbers is most often empirical and at haphazard. Furthermore, there is the danger that the assumptions made to accommodate efficiency are not valid for the system in question, resulting in loss of important information. This methodology is demonstrated by a number of papers by Pandit et al. [34, 35, 36], Wu et al. [50] and Eman [13] where the emphasis is on detection of chatter and tool wear.

The third methodology is based on extensive prior analysis and modeling of the system properties with the aim of including as much as possible of the resulting insights into the analysis and interpretation of the data. This requires the most effort but also provides the most reliable and easily interpreted results. Some parameters may not be known exactly, but usually limits and even a probability density may be specified which conveys knowledge about the machine or lack thereof. The modeling and analysis which this method calls for also pays off in better understanding of the machine's strengths and weaknesses for consideration during modifications or redesign. A number of researchers has adopted this methodology in their work. Braun [7] related the vibrations of a rolling element bearing to the elastic and dynamic properties of the races and the rolling elements in order to aid interpretation of vibration data. Diei and Dornfeld [10, 11] and Kannatey-Asibu et al. [22] have related the process variables of metal cutting to acoustic emissions from the process. Linear time-invariant systems are often sufficient as models of mechanical devices. Such models dominate in control theory and estimation and have also found application in machine monitoring. They are most commonly used to filter out all of the system's dynamics, leaving only an unknown stochastic part. If changes occur in the plant, a difference emerges between the plant and the model, resulting in additional signals mixed in with the stochastic residues. Willsky [48], Isermann [21] and Gertler [15] have published survey papers on these methods.

4.3.2 Data Processing Methods

Figure (4.1) shows the most typical of monitoring system hardware arrangements. The sensor may be of any kind — from oil particle detector to an optical surface finish transducer — however, in rotating machinery it is by far most common to sense vibrations. After some analog filtering (such as low pass filtering to prevent aliasing) the signal is sampled and stored in the computer for further processing. Initial processing may include digital filtering or any other operation from the big bag of signal processing tricks to enhance certain features of the signal or remove unwanted components. What happens next is that the signal is transformed in some way to reduce the data to a smaller set of numbers that is known to depend on the faults or other abnormal phenomena that the monitoring system is supposed to detect. This transformation may be of any conceivable kind and a great variety has been reported in the literature. In some cases, only an indication of some abnormality is found without further specification, calling for a second phase of diagnosis to follow the detection. Since the methods by which the data is reduced are at the heart of such monitoring systems, it is proper to describe selected examples in some detail.

The examples selected represent four categories into which most methods fall:

1. *Amplitude domain:* Measurements are characterized by some function of their amplitudes, such as their mean or variance.
2. *Time domain:* The time history of the measurements is used to identify such effects as delays, echos, trigger responses etc.
3. *Frequency domain:* The spectrum of the data is used, for example, to reveal periodicities or transfer characteristics.
4. *Model based methods:* Many methods use an analytically derived model in some form of the system as a reference to compare with actual observations. Analysis is based on differences between model predictions and observations.

Methods in the last category usually require most processing effort and previous analysis but they also provide results that are more exact and easily interpreted so long as the assumptions inherent to model are valid. Other methods are often less mathematical in nature but more empirical or heuristic and therefore also more specific to certain mechanical processes. Here, no attempt will be made to give a complete overview of existing methods but rather to introduce selected samples in order to illustrate the most popular methods. Other authors have written about mathematical tools useful for data analysis (Pau [37]), literature surveys on mechanical signature analysis (Volin [45] and Hundal [20]) and about machinery noise generation and propagation (Lyon [29]).

Amplitude Domain Methods

The simplest methods fall into this category, such as calculation of the mean and variance or other relevant parameters that characterize the distribution of the samples over the different amplitudes. It is also very easy to compile a distribution function based on observations which would reveal changes in the incoming signal. When, for example, a sinusoidal signal is added to one that normally contains only Gaussian noise, the bell-shaped distribution is deformed with "shoulders" appearing on both sides at a distance from the middle that equals the amplitude of the sine. The height of the shoulders indicates the power of the sinusoidal signal relative to the Gaussian noise (see [2] for more details). To test for the presence of the additional signal a χ^2 test might be performed.

Severity of machine vibrations is often characterized by the peak or the root mean square of vibration velocities but other measures in the amplitude domain have also been considered. Cempel [9] has investigated the application of certain dimensionless numbers to the evaluation of vibration signals from bearings and gears. These numbers (n_{ij}) are ratios of different moments of the density function $p(u)$ for the measurement amplitudes:

$$n_{ij} = \frac{\left[\int_{-\infty}^{\infty} |u|^i p(u) du\right]^{1/i}}{\left[\int_{-\infty}^{\infty} |u|^j p(u) du\right]^{1/j}} \tag{4.1}$$

This, for example, becomes equal to the fourth root of the Kurtosis for $i = 4$ and $j = 2$ but the Kurtosis is widely used to quantify the flatness of a distribution relative to a normal distribution. This is due to the high weight given to outlying samples by the fourth moment relative to the variance.

Dimensionless indicators like these have the advantage that they are largely independent on operating conditions such as load and speed so long as the machine operates in a similar fashion.

Time Domain Methods

It is often useful to examine the time history of a response to some trigger or impulse but the measurement of the response may be contaminated by noise or even reverberations of the response itself. To eliminate the noise a well know technique called time domain averaging is often applied especially when the noise contains power in the same frequency band as the response and filtering could therefore not eliminate the noise without distorting the response. The trick is to add up many measured responses which begin at the same reference time (e.g. the trigger time) and let the incoherent noise components cancel each other while the response component builds up. Let $s_i(t) = r(t) + n_i(t)$ be the measured signal composed of the deterministic response $r(t)$ and stochastic noise $n_i(t)$. Further let the noise be such that

$$E\{n_i(t)\} = 0 \quad \text{and} \quad E\{n_i(t)\, n_j(t)\} = \sigma_n(t)\delta_{ij} \tag{4.2}$$

Then we have

$$\mathrm{Var}\left\{\frac{1}{N}\sum_{i=1}^{N}(s_i(t) - r(t))\right\} = \frac{1}{N}\mathrm{Var}\{\,(s_i(t) - r(t))\,\} = \frac{\sigma_n^2(t)}{N} \tag{4.3}$$

Braun [6, 5] has described the application of this principle to the extraction of periodic signal components from rotating machinery. In that case the time series is broken up into segments that include one period and then the segments are averaged. Another interesting application is in the measurement of evoked potentials on the scalp, generated by electrical activity in the brain. Electrodes are mounted on the subjects head and his nervous system stimulated by audible clicks or flashes of light. The electrical response is severely contaminated by unrelated electrical activity such as normal brain waves and therefore the stimulus is repeated numerous times and the measurements added to extract the evoked response. The neurologist can then visually associate the shape of the response time history with functional aspects of the neural path between the sensory organs and the brain.

Frequency Domain Methods

Examining the frequency content of signals generated by rotating machinery is particularly useful since the individual shafts and gears or other moving components move in a deterministic and often repetitive fashion and power observed at certain frequencies can therefore easily be associated with some particular part of the machine. It has therefore been found very practical to calculate the spectrum of the signal and due to the importance of similar analysis in many other fileds a colorful flora of methods for spectral analysis has grown over the years (see Kay and Marple [23] and Kay [24]). These methods fall into two categories, parametric and nonparametric. The distinction lies in the format of the result of the calculations and the assumptions on which they are based. Nonparametric methods provide estimates of power directly at predetermined frequencies as in the case of the periodogram which also is most significant in that category. The parametric methods assume a model of the signal (and hence its spectrum) with a few degrees of freedom, thus reducing the problem to the one of estimating the value of the associated parameters. Whichever method is chosen, from either category, it can not be overemphasized that the key to successful estimation and meaningful interpretation of the results is that the assumptions underlying the calculations be consistent with the nature of the signal and its origin. The tutorial paper by Kay and Marple [23] contains a striking example of how different spectrum estimates come from applying the various methods to exactly the same data.

The periodogram is a proven and popular way to estimate spectra and efficient algorithms exist to calculate it. The fact that it is nonparametric represents both a strength and a weakness since no assumptions are implicitly made about the structure or other properties of the signal. Making no such assumptions allows for generality on one hand but since the result consists of half as many numbers as the original data, the variance of the result is substantial. Another disadvantage is that the periodogram calculates the spectrum as if the signal consisted of a finite number of sinusoids, equally spaced in frequency. This, of course, is almost never true of the signal but turns out to be a practical simplification with the price paid in biased estimates and "leakage" between adjacent frequency bins.

The periodogram is by far the most commonly applied data analysis method in machine monitoring practice and special devices are on the market that automatically monitor machinery by examining periodograms of vibrations or other measurements. Researchers have studied the applicability of parametric spectrum estimation to machine monitoring although such methods have not found their way to the factory floor much yet.

Most parametric spectrum estimation methods employ the same model of the signal, namely one that assumes that the signal is statistically equivalent to one generated by passing white noise through a linear difference equation such that each sample is a linear combination of current and earlier inputs and also earlier samples (outputs). These models are referred to as ARMA models where AR stands for autoregression or that part which depends on earlier samples and MA stands for moving average of the inputs. This is described mathematically by the difference equation

$$y_k = -\sum_{i=1}^{p} a_i y_{k-i} + \sum_{i=0}^{q} b_i w_{k-i} \tag{4.4}$$

where y_k is the current sample, w is the white noise of unity variance, a_i are parameters of the autoregression and b_i are parameters of the moving average. In the z domain this becomes

$$y(z) = \frac{B(z)}{A(z)} w(z) \tag{4.5}$$

with

$$A(z) = \sum_{i=0}^{p} a_i z^{-i} \text{ and } B(z) = \sum_{i=0}^{q} b_i z^{-i} \tag{4.6}$$

Given the parameters, the spectrum can be calculated by letting z take values on the unit circle:

$$S_{yy}(\omega) = \left| \frac{B(e^{-j\omega T})}{A(e^{-j\omega T})} \right|^2 \tag{4.7}$$

where T is the sampling interval.

It should be noted that the value of the white noise is not available for the estimation of the parameters and the problem is therefore not one of system identification.

A variety of methods exists to calculate the parameters from measurement data. Most employ some kind of least squares minimization of prediction error and many are specially constructed for computational efficiency. An additional problem associated with these methods is the determination of model order, if it is not known beforehand. For this, no absolute rules exist but some approximate criteria have been developed. It also turns out that the calculation of the parameters of a model that contains both an autoregression and a moving average is substantially more involved than if only either part is used. However, the spectral characteristics of an ARMA model can be approximated by using a longer sequence of either AR or MA. Using an AR model is especially popular, partly because of the AR model's ability to represent high peaks in the spectrum. This often leads to the use of very high order models, rendering it's advantage over a simple periodogram doubtful since data reduction is not much better.

In the context of machine monitoring it is especially important to point out some special characteristics that signals acquired from rotating machinery have. Since the excitation is predominantly due to the rotations of gears and shafts, the signals are going to contain strong periodic components which in the spectrum appear as a fundamental and a series of harmonics corresponding to the signal's Fourier decomposition. Let measurement y_k contain a deterministic part s_k and noise w_k:

$$y_k = s_k + w_k = \sum_{i=1}^{N} A_i \sin(\omega_i k + \phi_i) + w_k \tag{4.8}$$

where A_i, ω_i and ϕ_i are amplitude, frequency and phase of each of the N sinusoidal components, respectively. Properties of the sine function make it possible to write the deterministic part as an autoregression:

$$s_k = -\sum_{i=1}^{2N} a_i s_{k-i} \tag{4.9}$$

or equivalently:

$$y_k = -\sum_{i=1}^{2N} a_i y_{k-i} + \sum_{i=0}^{2N} a_i w_{k-i} \tag{4.10}$$

which is a degenerate form of an ARMA model where the coefficients of the autoregression and the moving average are the same. It can also be shown that the roots of the corresponding polynomials in the z-transform variable have are all located on the unit circle (see Ulrych and Clayton [42] and Stoica and Nehorai [41]). This shows that periodic signals with additive noise can not be modeled with AR or MA models only and further analysis shows that estimates based on such models will have biases in frequency and amplitude of the components. Many methods have been devised to go around this problem by using the degenerate ARMA form directly (see f. ex. Nehorai [32]) but most make use of an eigenequation that arises when the additive noise is white (see Pisarenko [38] and Vaccaro [43]). Although estimates based on this eigenequation are unbiased and capable of high accuracy under favorable conditions, the procedure suffers from the restriction to white noise, computational complexity and sensitivity to noise, requiring high signal to noise ratio to achieve good performance. Phase information is lost.

Cepstrum analysis is a frequency domain method that has proven useful to separate a primary signal from its echoes (Oppenheim and Schafer [33]). The cepstrum is defined as the logarithm of the Fourier transform of the signal:

$$\mathcal{C}(t^*) = \mathcal{F}^{-1} \log\left(\mathcal{F}(y(t))\right) \tag{4.11}$$

where the t^* domain is somewhat akin to the time domain, called quefrency. In this domain, the primary signal appears at low t^* values with the echoes appearing as periodical repetitions, separated in t^*. By keeping only the first instance (the primary) and reversing the calculations (inverse cepstrum), the original signal comes out without the echos. Lyon and Ordubadi [28] have reported an application of this procedure to signals from combustion engines.

State Space Model Methods

Linear state space methods are popular among control engineers for multi-input, multi-output controller design and state estimation. Consequently, many mechanical systems have been modeled using those techniques. Furthermore, the Kalman filter has become very popular for tracking and estimation of process variables. It is therefore not surprising that many monitoring systems attempt to use the information coded into such models and at the same time make use of the convenient structure that state space methods offer.

Let a linear or linearized system be modeled in the usual manner (see f. ex. Bryson and Ho [8]) by the following state-space equations:

$$\begin{aligned} x_{k+1} &= \Phi_k x_k + \Gamma_k u_k + w_k \qquad (4.12) \\ z_k &= H_k x_k + v_k \qquad (4.13) \end{aligned}$$

where x_k is the system state vector, Φ_k is the propagation matrix, Γ_k is the input matrix, H_k is observation matrix, z_k is vector of measurements and v_k and w_k are zero mean white noise vectors with covariances

$$E\{v_i v_j^T\} = R\delta_{ij} \qquad \text{and} \qquad E\{w_i w_j^T\} = Q\delta_{ij} \qquad (4.14)$$

where δ_{ij} is the Kronecker delta. The Kalman filter provides state estimates that are optimal in the sense of mean square estimation error by the equations

$$\begin{aligned} \hat{x}_{k+1|k} &= \Phi_k \hat{x}_{k|k} + \Gamma_k u_k \qquad (4.15) \\ \hat{x}_{k|k} &= \hat{x}_{k|k-1} + K_k \nu_k \qquad (4.16) \\ \hat{z}_k &= H_k \hat{x}_{k|k-1} \qquad (4.17) \end{aligned}$$

where K_k is the Kalman filter gain and ν_k is sequence of residuals. The residual sequence is generated by

$$\nu_k = z_k - \hat{z}_{k|k-1} \qquad (4.18)$$

Detection of changes in the system is divided into three phases, an alarm indicating that *some* change has occurred, identification of *which* change it was and then the estimation of its magnitude [48]. Most methods are based on analysis of the residuals sequence. Mehra and Peschon [31] suggested a simple technique for generating alarm. If the model reflects the system in normal operation and sufficiently accounts for all the system's dynamics, then the residual sequence should be Gaussian white noise assuming that stochastic inputs are also Gaussian. If a failure modifies the dynamics of the system, the system and its model will no longer match and the residual sequence will no longer be white. Testing for whiteness of the residuals therefore is equivalent to testing for the presence of a failure. A number of tests is suggested in [31]. For example, whiteness is checked by estimating autocorrelation lags of the residuals. All except the zeroeth lag should have zero mean and known covariance, which is simple to test. Also, statistics are given to test independence between components of the residuals vector, its mean and covariance.

Performing such simple tests on the innovations provides only indication of deviation from the model, but no information on the nature of the deviation. One solution is the multifilter method, which employs multiple Kalman filters, each based on a model of the system with one failure included in addition to one that corresponds to normal operation. The model that corresponds to the filter that produces the whitest residuals is assumed to be valid and thus the failure is identified. Another way to implement failure identification has been developed by Willsky and Jones [49] and Hall [16]. They have suggested ways to correlate the innovations to predetermined signatures of anticipated failures. An optimal test, based on the generalized likelihood ratio, is then performed to decide whether any of the signatures have appeared.

The above methods all suffer from sensitivity to modeling errors, because *all* the plant's dynamics must be included in the model and no errors in the model can be allowed if the nominal residues are to be white (see Kerr [25]).

4.3.3 Discussion

Monitoring systems that fit the description above represent the state of the art, with many successful applications. Critical review of these does however reveal a number of deficiencies that must be corrected for improved monitoring system performance and indeed, many improvements seem imminently possible.

- Many data processing strategies are based on heuristic indications when analysis and modeling of the mechanical process can yield much greater insights.

- The methods of the previous section can accommodate multiple sensors only if they provide inputs into a multi–input state–space model of the mechanical process. More flexibility is needed to accommodate information from diverse sources.

- When state–space models are used, the model must be correct and complete if the strategy is based on whiteness of residuals.

- Most methods fail to take into account information about the reliability of the results of the processing. For example, bias and variance in FFT based spectrum estimation is seldom considered.

- Prior expectations of what the result is likely to be are not considered.

- In many cases, the data processing strategy provides unnecessary estimates of known numbers. For example, a spectrum estimate of the motions of a rotating shaft will indicate the speed of rotation and proof of the fact that harmonics are equally spaced in frequency, neither of which it needs to tell us. Besides, spectral content between periodic components is often of no interest.

- Most methods require that a set of "failures" be specified beforehand and the monitoring system is designed around these. Other failures that may be too freakish to warrant inclusion may not be detected or be erroneously identified.

- The alarm oriented methods provide an indication of some failure or abnormality, without indicating which one, with minimal dela,· after the occurrence. Although indeed there are some systems where such swift detection is crucial, it is useful to keep in mind that processing times are getting shorter with faster computers. Also, necessary data accumulation before a reliable test statistic for alarm can be formed presents a fundamental lower limit to response time. It therefore seems reasonable to set this emphasis on fast detection aside if necessary to alleviate the more fundamental problems listed above.

In the monitoring system that has been developed in this research these problems are eliminated by its design.

4.4 NEW MONITORING SYSTEM ARCHITECTURE

Most monitoring systems that have been reporied in the literature so far have concentrated on the task of identifying malfunctions in machinery. In this research a more extensive role for the monitoring system is assumed by identifying three areas where the system could make important contributions:

- *Performance evaluation:* The monitoring system should not only report unacceptable performance but rather report the performance of the machine continuously during operation so that performance bottlenecks may be identified. This is likely to shed light on problems with quality, rejection rate and maintenance.

- *Design feedback:* Identification of problems with performance, efficiency and maintenance is valuable for the designer as a guide in design modifications or redesign.

- *Failure identification:* It remains an important task to identify quickly and reliably any malfunction that compromises safety and efficiency of the machine.

This effort to expand the role of the monitoring system both requires and suggests ways to alleviate some of the restrictions and problems of established systems as identified earlier. Instead of designing the monitoring system around a predefined and finite set of malfunctions, we need to be able to evaluate the machine's performance in more general terms. It is also desirable to incorporate as much information as possible into our assessment of the machine's performance. A natural way to accomplish this is to describe the machine and its performance in the designer's terms, i. e. the numerical values he would use to quantify the state of the machine and its performance. If this were realized and the operator could have reliable estimates of all design parameters that are unknown, subject to change or not directly controllable, the usual division into three phases, alarm, identification and magnitude estimation would go away and the operating state of the machine would be immediately clear. Another advantage is that prior information from the designer or that acquired through operating experience is expressed in these same terms.

The principal problem in realizing a monitoring scheme with the above capabilities is therefore that of estimating physically meaningful parameters with the inclusion of other information.

Let us now examine the resources that a monitoring system must have at its disposal to achieve the goals stated above. The vehicle for execution and orchestration of the system's activities is the choice of the architect. This may be conventional computer software or something more sophisticated, such as an expert system. As indicated in figure (4.2), the monitoring system uses information of three kinds, procedural information on how both machine and monitoring system operate, prior information about processes in the machine and measurement results. All these are integrated to extract information as reliably as possible on the performance of the machine.

4.4.1 Integration of Models and Measurements

The problem of integrating results from diverse sources has received considerable attention lately, especially in the fields of mobile robots and military systems. In both fields extensive use is made of a variety of dissimilar sensors to gather information (Luo and Kay [27]). For example, a mobile robot might use both sonar and stereo imaging data to get a better estimate of the distance to the nearest wall or obstacle than either one of the sources could provide. A corresponding example from machine monitoring is easy to come up with. Let us say that the stiffness of a structural member is being examined. Relevant information available may include:

- A recent direct measurement of the stiffness under conditions that have since changed, with an indication of the measurement error.
- Observation of the frequency of a vibration mode that is known to correspond to the member in question.
- Measurement of surface roughness that may have resulted from deflections of the member during the machining process.
- A model of how different operating conditions affect the stiffness of the member and how stiffness affects surface roughness.

All these should be integrated to provide a good estimate of the stiffness.

The probability density function (pdf) is a convenient format to convey knowledge about measurement or model projections. It also serves as an excellent vehicle for integration of information from diverse sources (see Durrant-Whyte for robot sensor application [12]) by using the rules and methods of standard probabilistic analysis. The pdf has the following advantages as format for machine monitoring information:

- The pdf carries in a natural way, not only the expected or most likely value of the result but also a measure of the value's reliability through its spread.
- The information can be specified in well defined modules in a compact form that is simple and intuitive.
- Conditional dependencies and correlations are easily accounted for.
- Systematic manipulation of probabilistic information is well established, for example by way of influence diagrams (Shachter [40]) which have been suggested as a framework for an expert system in a machine monitoring application (Agogino and Russell [1]).
- The theory of hypothesis testing and decision analysis is directly applicable to achieve best classification of the results and recommendations for subsequent actions

The numerical value of parameters may be estimated in a number of ways that are well known from estimation theory. Maximum likelihood estimation for example, finds the value of the parameter which makes the observations most probable, given a prior probability density of the observations but without considering prior knowledge or expectations of what the result will be. A more appealing estimation procedure for the monitoring application is "maximum a posteriori" (MAP) estimation, which combines the prior probability that is conveyed by the model with information on the reliability of the measurement (in terms of probability) and what the parameter values are expected to be, to arrive at a posterior density for the observation. The estimate is then selected as the value that maximizes this posterior density. Integration of model and measurement information is therefore inherent to the MAP estimation procedure.

4.4.2 Estimation Procedure

To realize as accurately as possible the condition of the machine, it is necessary to make the most of available information. It is assumed here that information available is of three kinds:

1. A model that predicts the outcome of measurements or features of the measurements, based on understanding of mechanical processes and uncertainties involved.

2. Measurements of relevant phenomena in the machine and any other quantities derived from the measurements, such as spectra.

3. Expectation of what the values of the physical parameters are likely to be and a measure of uncertainty about this expectation.

It is assumed that models contain parameters that have a clear physical interpretation, defined by the underlying analysis. Some of these parameters are not expected to change, such as the geometry of rigid bodies. Others are uncertain and may carry important information on the condition of the machine. In spite of the uncertainty about the value of these parameters, some values will be known to be more probable than others, by insight or experience. In other words, probability distributions for these parameters may be expected to be known, at least approximately.

Thorough understanding of the machine's function, especially with the aid of models describing the different processes will help identify any measurements or derived quantities that carry information on the parameter values. The integration of available information must therefore be able to take into account any number of dissimilar indicators that reflect on the parameters along with models and prior information on parameter values.

In most cases, parameters are not directly measurable but must instead be inferred from particular features of the measurement data set, that are known to be strongly dependent on those parameters. In the field of pattern recognition, the same situation arises and many of the concepts and methods established in that field are directly applicable here (see f. ex. Ho and Agrawal [19]). Let y be a vector of measurements and x be a vector of numbers that are sensitive to the unknown parameters with a mapping available that relates the two:

$$x = \phi(y) \tag{4.19}$$

In pattern recognition, this mapping is sometimes referred to as ***feature extraction*** and the vector x is known as ***pattern vector***. Using the model and knowledge of the nature of the features, it should be possible to construct another mapping from assumed values of the unknown parameters (vector θ) and the expected value of the pattern vector:

$$\bar{x} = E\{x\} = m(\theta) \tag{4.20}$$

The problem of feature extraction has no general solution. The feature extraction mapping must be tailormade for each case, using knowledge of the effect that changes in the wanted parameter have on the measurements. A myriad of signal processing and numerical analysis techniques exist that may be used in construction of $\phi(y)$. The other mapping $m(\theta)$ is on the other hand more likely to be directly extractable from the model.

Now the problem is to combine the information from the three sources into an estimate of θ. This is accomplished by a well known procedure from estimation theory called "maximum a posteriori" (MAP) estimation which is based on maximization of the posterior probability density for θ given the measured pattern vector x (see f. ex. van Trees [44]). With the probability densities $p(x|\bar{x}) = p(x|\theta)$, $p(\theta)$ and $p(x)$ it is easy to calculate the posterior density:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \tag{4.21}$$

and the estimate is specified as:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta|x) \tag{4.22}$$

For continuous probability densities, this maximum may be found by gradient search methods or other nonlinear optimization techniques. For some densities, such as the Gaussian it is easier to work with the logarithm of the posterior density; this does not affect the location of the maximum. It is also noteworthy that the prior density of x can be calculated from the other two densities given:

$$p(x) = \int p(x|\theta)p(\theta)d\theta \tag{4.23}$$

However, this is not needed to locate the maximum as it does not depend on θ.

It is important to note that in the feature extraction mapping, $x = \phi(y)$, the original data y can be gathered from many different sources and the elements of x may have any meaning or none at all, only so long as it depends on the parameters that are being estimated and the dependency is to some extent understood. Since in all the ensuing analysis, x is referenced only in terms of the probability of its occurrence, there are no limitations on what can be used to make up the pattern. Herein lies the power of this approach.

It is useful at this point to establish some measure of the quality of the estimate. A measure commonly used in estimation theory is the Fisher information matrix which for the posterior density of θ is defined as:

$$J_{\theta|x} = E\left\{\nabla_{\theta}[\ln p(\theta|x)]\ \nabla_{\theta}^{T}[\ln p(\theta|x)]\right\} \tag{4.24}$$

This matrix establishes lower bounds on covariances of the MAP estimate, in particular:

$$\mathrm{Var}\{\hat{\theta}_{MAP,i}\} \geq [J^{-1}]_{ii} \tag{4.25}$$

In other words, the diagonal elements in the inverse of the information matrix are lower bounds on the variance of the elements of the parameter vector estimate.

Substituting (4.21) into the above and assuming that estimates are unbiased and that the error in the pattern vector is independent of θ, we get:

$$J_{\theta|x} = E\left\{\nabla_\theta[\ln(p(x|\theta) + \ln p(\theta))]\ \nabla_\theta^T[\ln(p(x|\theta) + \ln p(\theta))]\right\}$$

$$= E\left\{\nabla_\theta[\ln p(x|\theta)]\ \nabla_\theta^T[\ln p(x|\theta)]\right\} + E\left\{\nabla_\theta[\ln p(\theta)]\ \nabla_\theta^T[\ln p(\theta)]\right\} = J_{x|\theta} + J_\theta \tag{4.26}$$

Let us now examine a simple but useful example, where a scalar parameter, θ, is being estimated from a scalar pattern with both having a Gaussian probability density:

$$p(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{(x - m(\theta))^2}{2\sigma_x^2}\right) \quad \text{i.e.} \quad N(m(\theta), \sigma_x^2)$$

$$p(\theta) = \frac{1}{\sqrt{2\pi}\sigma_\theta} \exp\left(-\frac{(\theta - \theta^*)^2}{2\sigma_\theta^2}\right) \quad \text{i.e.} \quad N(\theta^*, \sigma_\theta^2)$$

using the logarithm of the posterior density, we have:

$$\hat{\theta}_{MAP} = \arg\max_\theta p(\theta|x) = \arg\max_\theta [\ln p(x|\theta) + \ln p(\theta)] \tag{4.27}$$

or

$$\frac{\partial}{\partial\theta} \ln p(x|\theta) + \frac{\partial}{\partial\theta} \ln p(\theta)\Big|_{\theta=\hat{\theta}_{MAP}} = 0 \tag{4.28}$$

and substituting the Gaussian densities, this becomes:

$$\frac{1}{\sigma_x^2}(x - m(\theta))\frac{\partial m}{\partial\theta} - \frac{1}{\sigma_\theta^2}(\theta - \bar{\theta})\Big|_{\theta=\hat{\theta}_{MAP}} = 0 \tag{4.29}$$

This can not be reduced much further without knowledge of $m(\theta)$ but if it is also assumed that the expected value of the pattern is proportional to the parameter, $m(\theta) = \alpha\theta$ we have:

$$\frac{1}{\sigma_x^2}(x - \alpha\hat{\theta}_{MAP})\alpha - \frac{1}{\sigma_\theta^2}(\hat{\theta}_{MAP} - \theta^*) \tag{4.30}$$

which yields the estimate:

$$\hat{\theta}_{MAP} = \frac{\theta^*\sigma_x^2 + x\alpha\sigma_\theta^2}{\sigma_x^2 + \alpha^2\sigma_\theta^2} \tag{4.31}$$

This shows that if little is known about the value of the parameter (σ_θ is large) then the estimate is going to be close to x/α which is where the model wants it to be to match the pattern. If on the other hand there is good prior certainty about the value of the parameter and either high variance in the pattern or little sensitivity to the parameter, the estimate is going to lean more towards the prior value. This is especially effective when many parameters are needed to match a complex pattern because it conveys information on which parameters are most free to be adjusted and which ones should be kept close to the prior expected value. A similar weighting of prior and posterior information happens in the Kalman filter, where a compromise is reached between model predictions of the state vector and measurement results, based on the relative magnitude of the covariances involved.

The Fisher information for the result above is also easily calculated:

$$J_{x|\theta} = E\left\{\left(\frac{\partial}{\partial\theta}\ln p(x|\theta)\right)^2\right\} = \frac{1}{\sigma_x^4}\left(\frac{\partial m}{\partial\theta}\right)^2 E\left\{(x - m(x))^2\right\} = \frac{1}{\sigma_x^2}\left(\frac{\partial m}{\partial\theta}\right)^2 \tag{4.32}$$

and

$$J_\theta = E\left\{\left(\frac{\partial}{\partial\theta}\ln p(\theta)\right)^2\right\} = E\left\{\left(\frac{1}{\sigma_\theta^2}(\theta - \theta^*)\right)^2\right\} = \frac{1}{\sigma_\theta^2} \tag{4.33}$$

and hence:

$$J_{\theta|x} = \frac{1}{\sigma_x^2}\left(\frac{\partial m}{\partial\theta}\right)^2 + \frac{1}{\sigma_\theta^2} \geq \frac{1}{\sigma_\theta^2} \tag{4.34}$$

This indicates that tight distributions and good sensitivity of the pattern to the parameter are necessary for good estimates.

Since the feature extraction mapping may often be quite complex, it is probably a useful approximation in practice to assume that the prior density of the pattern is Gaussian and that it is unbiased by design. The variance of this distribution may be determined by simulation or analysis. On the other hand, the prior density for θ may have any conceivable shape. However, it is convenient whenever possible to assume a Gaussian density for the parameter vector also, because as before, the problem becomes more tractable. Let us develop the all Gaussian case further with numerical solution in mind.

Let a pattern vector x be available and a model vector $m(\theta)$ which represents our a priori expectation of what the value of x will be, where θ is the vector of unknown parameters. The probability density of x is therefore:

$$p(x\,|\,\theta) = (2\pi)^{-n/2}\,|\Sigma_x|^{-1/2}\exp\left\{-\frac{1}{2}(x - m(\theta))^T\Sigma_x^{-1}(x - m(\theta))\right\} \tag{4.35}$$

where n is the dimension of the pattern vector, x. In a similar fashion the density of the parameter vector θ is:

$$p(\theta) = (2\pi)^{-k/2}\,|\Sigma_\theta|^{-1/2}\exp\left\{-\frac{1}{2}(\theta - \theta^*)^T\Sigma_\theta^{-1}(\theta - \theta^*)\right\} \tag{4.36}$$

where k is the dimension of the parameter vector, θ and θ^* is the prior expected value of the parameters. As shown earlier, it is helpful for the analysis to take logarithm of the posterior density but it also helps convergence of gradient based algorithms, because if a pattern or parameter is more than, say three standard deviations away from the mean, the slope of the density is small and the solution will converge slowly. This is obviously not the case with the logarithm of the density, which in the Gaussian case is a parabola. Another problem often encountered in nonlinear optimization is the great diversity of numerical values that the different parameters are likely to assume. A good example is provided by hydrostatic bearings, where parameters range from 5×10^{-5} for bearing gaps in meters to 1×10^9 for the bulk modulus of the fluid, in Pascals. In the optimization, this leads to a very elongated surface that is searched for extrema. Some gradient based methods, such as the Newton method, take into account the curvature of the surface and are thus able to compensate for this effect at the cost of more calculations. Since in this case we are dealing with Gaussian density functions, it is also possible to scale the parameters with their standard deviation and perform the maximization with respect to the scaled parameter. Then the problem is:

$$\hat{t}_{MAP} = \arg\max_t \; p(x\,|\,t)\,p(t) \tag{4.37}$$

where

$$t = \Sigma_\theta^{-1/2}(\theta - \theta^*) \tag{4.38}$$

and hence

$$\theta = \theta^* + \Sigma_\theta^{1/2} t \tag{4.39}$$

with probability density

$$p(t) = (2\pi)^{-k/2} \exp\left\{-\frac{1}{2}t^T t\right\} \tag{4.40}$$

Then the gradient with respect to t is:

$$\nabla_t \ln\{p(x\,|\,t)p(t)\} = \nabla_t \ln p(x\,|\,t) + \nabla_t \ln p(t) =$$

$$= (x - m(\theta))^T \Sigma_x^{-1} \frac{\partial m(\theta)}{\partial \theta}\frac{\partial \theta}{\partial t} - t \tag{4.41}$$

where $\partial m(\theta)/\partial\theta$ is a Jacobian which normally would have to be evaluated numerically and $\partial\theta/\partial t$ is simply:

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial t}\left\{\theta^* + \Sigma_\theta^{1/2} t\right\} = \Sigma_\theta^{1/2} \tag{4.42}$$

In most cases, individual parameters (i.e. components of θ) can be treated as statistically independent and the covariance matrices Σ_θ and $\partial\theta/\partial t$ are therefore diagonal, rendering their inversion, rootfinding and multiplication quite trivial.

The simplest optimization algorithm is to calculate the gradient of the objective function, in this case using (4.41) and move the parameter vector in that direction:

$$t_{k+1} = t_k + s\nabla_t \ln\{p(x \mid t)p(t)\} \tag{4.43}$$

where s is a heuristic parameter that controls the stepsize. Algorithms of this type, known as steepest descent algorithms, are notorious for their slow convergence. A better algorithm is developed in a later section.

Since the Gaussian distribution is unimodal, false extrema can only arise through a deficiency in the feature selection that determines the pattern vector x. Careful selection along with thorough understanding of how the pattern depends on the parameters will minimize the risk of running into false extrema but no guaranteed general method can be given.

It is now time to consider the case when the probability density of a parameter can not realistically be assumed to be Gaussian. Two instances are of particular interest:

- The parameter is completely unknown except for upper and lower bounds.
- The prior density of the parameter (i.e. a component of the parameter vector) is multimodal with concentrations at two or more specific values.

The first case is easily treated by giving the parameter a Gaussian density with a very large variance and imposing hard limits on the parameter as the parameter is being adjusted during the gradient search procedure.

The second case arises naturally when a parameter reflects different states of the machine, assuming a different value for each state, with the exact numerical value being subject to some uncertainty. The prior density is therefore of the form:

$$p(\theta) = \sum_{i=1}^{N} \pi_i p_i(\theta) \tag{4.44}$$

where each of N distributions $p_i(\theta)$ is unimodal and the π_i are the prior probabilities of each machine state, i. e. the probability of the corresponding $p_i(\theta)$ being the relevant one. To find the parameter value with highest posterior probability, the maximization must be performed for each mode of the distribution and the posterior probability multiplied by the corresponding π_i and that value selected that gives the highest result. This way, a classification of the machine state is achieved along with the parameter estimate. This method may also be used if the prior density of the parameter is multimodal for other reasons by approximating the density with a weighted sum of Gaussian densities. In this case, the identity of the mode selected will probably not be as meaningful as before.

It has now been shown how a parameter can be estimated using both prior information and measurement data or other posterior indicators. The quality of this estimate has yet to be considered. In general, the posterior density is very difficult to derive analytically with numerical methods being the best bet. The Fisher information for vector valued x and θ is easily calculated in exactly the same fashion as before (equations 4.32 to 4.34):

$$J_{\theta|x} = \left(\frac{\partial m(\theta)}{\partial \theta}\right)^T \Sigma_x^{-1} \left(\frac{\partial m(\theta)}{\partial \theta}\right) + \Sigma_\theta^{-1} \tag{4.45}$$

The inverse of this matrix is a minimal covariance matrix, in the sense that the corresponding equal probability ellipsoids are all inside the real equal probability surfaces of the same value. If the model $m(\theta)$ is linear and prior densities are Gaussian, $J_{\theta|x}^{-1}$ is the exact posterior covariance matrix and the posterior density is Gaussian.

The eigenvectors and eigenvalues of the covariance matrix may help in identifying lost degrees of freedom in the parameter space of the model, for example due to overparametrization. If the posterior covariance in estimation of two parameters has eigenvectors close to say, $[1, -1]$ and $[1, 1]$, associated with a very high eigenvalue (close to the prior variance) and a very low one, respectively, it may be inferred that little information was retrieved from the pattern about the difference between the parameters but their sum was estimated accurately. This would indeed be the case if the model could be expressed in terms of a sum of the parameters or their quotient. This again may suggest modifications to the pattern that lead to improved estimation.

4.4.3 Sensor Data Processing and Fusion

In the machine monitoring system proposed here, it is the purpose of the sensor data processing system to sample the data from all the different sensors installed, as demanded by subsequent estimation procedures, process the data from each sensor as needed (e.g. filtering, spectral analysis etc.) and fuse the data from different sensors into one coherent pattern that is then forwarded to the estimation procedure. Let us take a look at some of the issues involved:

Data processing may employ any of the methods established in signal processing, numerical analysis and statistics to extract relevant features from data sets, such as spectral components from time series. In this research, the use of result variances is emphasized.

Sensor data fusion refers to the combination of data from multiple (possibly dissimilar) sensors into a single result. Many sensors may carry some information on a phenomenon of interest, which when combined makes one reliable result. In machine monitoring the fusion of sensor data happens normally at a low level since most information is gathered in the form of time series of accelerations, motions, forces or pressures as opposed to robotic and military systems where, for example objects are located by combining information from vision and sonar or radar systems.

Hardware requirements include the ability to acquire data from many sensors simultaneously at any time and complete processing without unreasonable delay.

Logical sensors is a concept introduced by Henderson and Shilcrat [18]. Its purpose is to make the the specification of a sensor abstract, leaving out the details of the device itself and subsequent processing. Thus the logical sensor 'senses' some feature of the data from a physical sensor or other logical sensors rather than the raw data itself. Such constructs facilitate automated reasoning about the fusion and processing of sensor data and may be useful for monitoring systems that employ such reasoning techniques.

4.4.4 Machine Models

In this context, the term model shall mean any prior information that may be supplied about expected behavior of the machine. This may be based on analysis of physical processes, earlier observations, statistics etc. It is important to make as much as possible of such prior information available to the monitoring system. The following issues should be considered when modeling the machine:

- To facilitate interpretation and integration with other information, the model should be derived in terms of actual physical parameters and quantities.
- It may not be necessary for a model to account for all aspects of all observations. For example, it may suffice to model dynamics of a process up to a certain frequency and ignore higher frequencies, assuming that no information about machine performance is contained in that portion.
- It is important to allow for lack of information where appropriate. This may for example be done by assigning probabilities to different states of processes that are too complex to be modeled or simply badly understood.
- The best model is the one that describes the subject adequately in as simple manner as possible.

It should also be noted that there is nothing wrong with empirical models, so long as they relate parameters being estimated to the observations. It is important though, to add any uncertainty in the model projections to the variance in the pattern to account for the missing information.

4.4.5 Procedural Information

The procedural information is twofold. One part describes the different operating modes of the machine such as operation sequences, conditions and prerequisites for operations. The other part has to do with the monitoring system itself, its modes of operation, strategies and programs. This information lends itself to incorporation into an expert system that may decide or suggest procedures to be taken in any situation that may arise during operation.

4.5 AN EFFICIENT ESTIMATION ALGORITHM

We have seen that a good estimate of the parameters is found by maximizing their posterior density:

$$\hat{\theta}_{MAP} = \arg \max_{\theta} p(x|\theta)p(\theta) \tag{4.46}$$

It has also been argued that it is better to work with the normalized parameters t and the logarithm of the posterior density:

$$\hat{t}_{MAP} = \arg \max_{t} [\ln p(x|t) + \ln p(t)] \tag{4.47}$$

yielding, of course, the same result. At the maximum, the gradient of the posterior density must be zero, which in the case of Gaussian prior densities gives (see equation (4.41)):

$$(x - m(\theta))^T \Sigma_x^{-1} \frac{\partial m(\theta)}{\partial \theta} \Sigma_\theta^{1/2} - t = 0 \tag{4.48}$$

Because of nonlinearities in $m(\theta)$ it is necessary to employ a numerical search algorithm. Steepest descent gradient algorithms of the type (4.43) are known to have slow convergence and in addition it may be expected that the effort to calculate the gradient of $m(\theta)$ needed at each step is nontrivial. One method to overcome this is to calculate the gradient at some initial point and then proceed in that direction until a maximum on that line has been reached. At that point, a new gradient is calculated and so forth. This saves gradient evaluations but still, a heuristic stepsize parameter must be used to control the convergence towards the intermediate maxima. A similar principle is used in the algorithm that is introduced here, however the special structure of the surface that is being searched for maximum is used to locate approximately the intermediate maxima in one step. Let θ_k (and hence also t_k) be the estimated parameter vector at step k with $m(\theta)$ and $\partial m(\theta)/\partial\theta$ being evaluated at this point. At the next parameter estimate, θ_{k+1}, the gradient with respect to t will be:

$$\nabla_t \log\{p(x\,|\,t)p(t)\}|_{k+1} = (x - m(\theta_{k+1}))^T \Sigma_x^{-1} \frac{\partial m(\theta_{k+1})}{\partial \theta} \Sigma_\theta^{1/2} - t_{k+1} \tag{4.49}$$

where θ_{k+1} and t_{k+1} are linearly related to each other:

$$\theta_{k+1} = \theta^* + \Sigma_\theta^{1/2} t_{k+1} \tag{4.50}$$

Using a first order approximation to $m(\theta)$ we have:

$$m(\theta_{k+1}) \approx m(\theta_k) + \frac{\partial m(\theta_k)}{\partial \theta}(\theta_{k+1} - \theta_k) \tag{4.51}$$

and

$$\frac{\partial m(\theta_{k+1})}{\partial \theta} \approx \frac{\partial m(\theta_k)}{\partial \theta} \tag{4.52}$$

and hence the gradient at the new location is:

$$\nabla_t \log\{p(x\,|\,t)p(t)\}|_{k+1} =$$

$$= (x - m(\theta_k) - \frac{\partial m(\theta_k)}{\partial \theta}(\theta_{k+1} - \theta_k))^T \Sigma_x^{-1} \frac{\partial m(\theta_k)}{\partial \theta} \Sigma_\theta^{1/2} - t_{k+1} \tag{4.53}$$

Equating this gradient to zero, we have:

$$b^T - \Sigma_\theta^{1/2} A^T \theta^* - \left\{\Sigma_\theta^{1/2} A \Sigma_\theta^{1/2} + I\right\}^T t_{k+1} = 0 \tag{4.54}$$

where

$$b = (x - m(\theta_k) + \frac{\partial m(\theta_k)}{\partial \theta}\theta_k)^T \Sigma_x^{-1} \frac{\partial m(\theta_k)}{\partial \theta} \Sigma_\theta^{1/2} \tag{4.55}$$

and

$$A = \left(\frac{\partial m(\theta_k)}{\partial \theta}\right)^T \Sigma_x^{-1} \left(\frac{\partial m(\theta_k)}{\partial \theta}\right) \tag{4.56}$$

Equation (4.54) above is efficiently solved for t_{k+1} by Gaussian elimination, without matrix inversion being necessary.

The algorithm is therefore compactly stated as follows:

```
let $t_0 = 0$ and $k = 0$
repeat:
    $\theta_k = \theta^* + \Sigma_\theta^{1/2} t_k$
    calculate $m(\theta_k)$ and $\partial m(\theta_k)/\partial \theta$
    calculate $t_{k+1}$ using (4.54)
until $(t_{k+1} - t_k)^T (t_{k+1} - t_k) < \epsilon$
```

In the termination rule, ϵ is some small number that determines the minimum length of the last step.

It is interesting and useful to note that the matrix in equation (4.54) is the Fisher information matrix with respect to t:

$$J_{t|x} = \Sigma_\theta^{1/2} A \Sigma_\theta^{1/2} + I \tag{4.57}$$

and hence, after the final step, k

$$J_{\theta|x} = \left(\frac{\partial m(\theta_k)}{\partial \theta}\right)^T \Sigma_x^{-1} \left(\frac{\partial m(\theta_k)}{\partial \theta}\right) + \Sigma_\theta^{-1} \tag{4.58}$$

These matrices are readily available when the algorithm has converged. It is also easy to see that if the model $m(\theta)$ is linear in the parameters, the Hessian of the surface is equal to the Fisher information matrix and the posterior density is Gaussian if all the prior densities are. Equation (4.54) is therefore similar to update equations used in Newton-type search algorithms except that the second derivative of the model is missing whereas the second derivatives of the densities (indicated by variances) are included. This accounts for the good convergence rate of the algorithm which, although it is not quadratic as for the Newton algorithms, is "superlinear" with convergence parameter usually below 0.1, i. e.

$$||t_{k+1} - \hat{t}_{MAP}|| < .1||t_k - \hat{t}_{MAP}|| \tag{4.59}$$

Figures 4.3, 4.4 and 4.5 show examples of convergence toward the maximum of the posterior density for a two dimensional pattern and two parameters (two parameters selected for easy visualization). In all cases prior densities of parameters and pattern are:

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 3 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) \tag{4.60}$$

and

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) \tag{4.61}$$

Figures 4.4 and 4.5 where the models are:

$$m(\theta) = \begin{bmatrix} \theta_1^3\theta_2 \\ \theta_1/\theta_2^3 \end{bmatrix} \qquad \text{and} \qquad m(\theta) = \begin{bmatrix} \theta_1^2\theta_2 \\ \theta_1^2/\theta_2^3 + \theta_2 \end{bmatrix} \tag{4.62}$$

respectively, were chosen because they showed unusually erratic convergence out of a few different nonlinear models tried. The iterations start out from the a priori estimate of the parameters (in this case $\theta = [1, 1]^T$) and converge to the maximum in 4–5 iterations. Figure (4.6) shows how the length of the error vector $||t_k - \hat{t}_{MAP}||$ changes in a geometric progression with convergence parameter of about .03 per iteration.

If the model $m(\theta)$ is linear in the parameter vector, the algorithm converges in one step. Figure (4.5) shows a degenerate surface where the model has only one degree of freedom. Observing the principal axis of the surface, it is clear that the sum of the parameters has a tight distribution, whereas the difference has a large variance. The surface does curve slightly along the long axis, corresponding to the (large) prior variance specified for each parameter.

4.6 CLASSIFICATION

Having gathered all relevant information and estimated relevant parameters, there are decisions that must be made on how to react. Here, we treat the case where a parameter describes a condition that must not exceed a certain limit. If the limit is exceeded, a predetermined procedure is performed to correct the problem. Otherwise, nothing needs to be done. We assume that the distribution of the MAP estimate is approximately known, either from simulation or by observing that the posterior density is close enough to Gaussian to allow the minimum variance provided by the Fisher information matrix to be used. Let us say that a parameter θ has been estimated and its distribution is approximately known to be as shown on fig. (4.7). Let also θ_{max} be the maximum allowable value of θ in the sense that if exceeded, the machine is not going to be able to meet specifications and corrections need to be made. With the distribution shown, there is probability β that the true value of θ is actually too high even though the estimate is under the limit. In this situation, it is intuitive that if the consequences of continued operation under faulty conditions are very serious when at the same time a check by the operator is quick and easy to do, it is wise to alert the operator even when β is very low. Conversely, an unimportant fault that requires long downtime to double-check, should call for higher β before action is recommended. To cast this into an explicit rule, we need to define some terms. Define the following costs:

C_D is the cost associated with correctly detecting faulty conditions, including downtime, repairs, etc.

C_M is the cost of incorrectly assuming no fault when in fact the machine is running under faulty conditions (miss).

C_{FA} is the cost of incorrectly assuming that a fault has occurred when it has, in fact, not (false alarm).

Having $\hat{\theta}_{MAP}$, θ_{max}, β and the above cost figures, we need to decide whether or not to assume $\theta_{real} > \theta_{max}$. With each possibility the decision can be right or wrong as shown on the simple decision tree in fig. (4.8). The expected cost associated with each situation can be simply found by multiplying the probability of occurrence (either β or $(1-\beta)$) with the relevant cost and the best decision is reasonably identified as the one with lowest expected cost. Hence, we should declare fault if:

$$\beta C_D + (1-\beta)C_{FA} < \beta C_M \qquad (4.63)$$

or

$$\beta > \frac{C_{FA}}{C_M + C_{FA} - C_D} \qquad (4.64)$$

Normally, the costs are such that $C_M > C_D > C_{FA}$ ensuring a positive denominator.

Example: The stiffness of a hydraulic actuator has been estimated at 100 MN/m and the standard deviation of that estimate (presumed Gaussian) is found to be about 10 MN/m. Experience has shown that if the stiffness falls below 75 MN/m, chatter occurs, which will destroy the workpiece if it is not corrected. If alarm is sounded, a direct measurement that reliably indicates the stiffness must be performed, halting the machine for 1/2 hour. If the alarm was false, machining can then continue, otherwise we expect it to take 1 hour to repair the actuator. If the fault goes unnoticed, it will take 10 hours to redo the piece plus the one and a half to measure and fix the actuator. Integrating the Gaussian density below 75 MN/m gives $\beta = .0062$ and since

$$.0062 < \frac{.5}{11.5 + .5 - 1.5} = .0476 \tag{4.65}$$

no alarm is sounded. In fact, fault should not be assumed until $\beta > .0476$ which with the given standard deviation occurs when the stiffness is estimated at 83.3 MN/m or less.

4.7 DISCUSSION OF METHODOLOGY

In precision engineering the attitude towards error analysis has been a subject of some consideration (see Evans [14]). The prevailing attitude is referred to as "determinism" meaning that errors should not be looked at as stochastic phenomena that defy understanding but rather as deterministic results, obeying engineering principles. At the same time, it is acknowledged that some phenomena are, in spite their deterministic nature, too complex for the currently available techniques of engineering analysis and must therefore, lacking any better approach, be looked at as stochastic. This attitude is, of course, valid in many other disciplines.

The monitoring methodology that is introduced here is fully in line with this attitude of determinism, despite all the probabilistic analysis. In this methodology, a careful separation of what is known and what not is required and the inclusion of all available information is encouraged. Implementation of a monitoring system of this kind must take into account some important issues:

- The probabilistic analysis presented earlier is based exclusively on Gaussian probability densities. This is justified by the following arguments:

  1. Very rarely will it be possible to specify the exact shape of the probability density of a pattern or a parameter. In specifying prior densities, the information provided to the monitoring system is likely to state that a parameter has a certain value, plus/minus some expected deviation. The Gaussian assumption is also standard in signal processing so that estimates of pattern variances are likely to be based on Gaussian noise assumptions that may or may not be valid. Obviously, no firm conclusions can be drawn about the exact density shapes. The moral is that the Gaussian densities play the role of *certainty indicators* rather than exact descriptions of probabilities. All

conclusions that rely directly on the shape of the Gaussian density, such as some of the classification procedures in the previous section, should be looked at as guidelines only, which may still be important.

2. Nonlinearities in the model will skew or otherwise deform the posterior density from the standard Gaussian shape. However, it takes a "very nonlinear" model to render the posterior density significantly non-Gaussian. Since the input densities were only approximately Gaussian anyway, it is reasonable to approximate again the posterior density in the same way.
3. By approximating the posterior with a Gaussian density it can become a new, updated prior density for a subsequent estimation.

- Some parameters may not be accurately known but guaranteed to remain at a fixed value throughout the operation. Such parameters can be estimated once and for all in the beginning, perhaps using special measurements and patterns for that purpose and then assumed to be known constants if the posterior variance becomes negligible.

- The decay of information is also important. If a parameter is expected to show slow changes, perhaps due to wear, heating or other gradual effects, it may be safe to keep the posterior density of that parameter as prior for the next estimation with some increase in variance (which depends on age of the estimate) so that it does gradually become "forgotten". Correspondingly, a fast changing parameter should be forgotten soon and the prior returned to its original specification.

- It is possible to set up special traps in the software to detect inconsistencies in estimates. For example, if a parameter is common to two patterns, and estimation based on each pattern gives a very different value relative to the variance, there must be something wrong with the information provided, for example bias in either pattern that is not accounted for in the corresponding model.

- It is wise to specify a generous variance in the prior densities of the parameters since it is also the maximum variance that can come out of the estimation (the pattern never reduces information). At least, all sources of error in the assumed value should be considered and added up with perhaps a supplement for safety in the spirit of good engineering pessimism.

- The fact that the designer of the monitoring system is *forced* to think about what is already known and how reliable the results are, should be considered a feature of this approach, not a drawback.

4.8 COMPUTER HARDWARE AND INSTRUMENTATION

The hardware for the monitoring system was selected to provide versatility during the system's development. All the selections have performed excellently and proven themselves well suited.

The computer hardware consists of two systems, connected by network. One is a multiprocessor, intended for real–time work and data processing. The other is a workstation which handles higher level processing, code development and user interface. Specifically the following items were selected:

1. VME bus standard for board level products, materialized by backplane and chassis from Electronic Solutions model "VME power cage PCV8020D-160V60". It can accommodate 20 boards and supply them with power.
2. Motorola MVME-133 processors, each with 1MB RAM. Three identical ones were selected to demonstrate utility of multiprocessing in a monitoring system.
3. Motorola MVME-225-2, 2 Megabyte dynamic RAM memory module. This memory resides on the bus and is equally accessible to all processors.
4. Motorola MVME-330 Ethernet controller for communication with disk server and workstation.
5. SUN Microsystems workstation, type 3/60. Used for all code development, supervisory control of monitoring system and for data presentation.

These selections benefit from existing facilities and experience acquired in other laboratories at Stanford.

The portion of the monitoring system that is developed in this research requires measurements of pressures in the hydraulic oil circuit and forces exerted on the cutting tool. For these measurements, piezoelectric sensors were selected for their small size, sensitivity and stiffness. These sensors are based on the piezoelectric effect by which electric charges arise in some crystals (quartz is most commonly used) when they are subject to stress in certain directions. Since the internal resistance of the crystal is finite, the charge will eventually leak and disappear, rendering the sensor unable to measure very low frequency or constant signals. Many sensors of this type have also electronics built into the sensor housing that converts the weak charge signal into a strong voltage or current signal that can be transmitted without distortion over long cables.

The following sensors and amplifiers were selected for the monitoring system. All are made by PCB Piezotronics Inc.:

- Pressure sensor (5), type 112A23 have a sensitivity of 7.25×10^{-6} V/Pa (50 mV/psi). Resolution is 28 Pa (.004 psi) and discharge time constant is 1 sec.

- Force sensor (1), type 208A02 has sensitivity of .011 V/N (.05 V/lbf). Resolution is .09 N (.02 lbf) and discharge time constant is 500 sec.
- Force sensor (1), type 209A has sensitivity of .5 V/N (2.2 V/lbf). Resolution is .2 mN and discharge time constant is 1 sec.
- Amplifier, type 483B07 has 12 channels with amplification range from 0 to 100.

The interface between the real-time computer and the instrumentation is furnished by a VME-bus analog to digital conversion card, type XVME-566 from Xycom. This card can accomodate up to 32 channels of single ended analog inputs multiplexed into one analog to digital conversion circuit. Word length is 12 bits and maximum sampling rate is 100 kHz. Special capabilities of the card include programmable channel sequencing, programmable gains and on board memory with circuitry for automatic sampling without processor intervention.

The anti-aliasing filters are of type RIFA–PBA 3257, designed for use in digital audio equipment. They are 10 pole linear phase elliptical type filters with break frequency at 7 kHz and 64 dB stopband rejection. One filter is used for each channel of the A/D converter. A sampling rate of 16 kHz is normally used.

Figure 4.9 shows schematically the connections between instrumentation, real-time computer and SUN-workstation.

4.9 COMPUTER SOFTWARE

Special software was written for the real-time system to carry out the data processing necessary for the monitoring system. The software was designed to make use of the fact that many data analysis procedures can be broken into steps that are standard operations, such as calculations of statistical moments, correlations and spectra. This can be used in the following ways:

1. A data processing "language" can be defined in which the programs are composed of the standard operations. This makes the construction of data processing programs much more efficient.
2. With a relatively simple mechanism and proper syntax of the program, it is possible to enable many processors to work on the same program.
3. Standard operations have well defined properties, opening the possibility of having an expert system select operations or even construct programs.

Intermediate and end results of the processing are available in variables that reside in memory common to the entire real-time system. Using specially constructed routines, the contents of these variables can be accessed by the MAT-

LAB [1] program which handles integration of information and presentation of results to the user.

The MATLAB program is in many ways well suited to handle the highest level of the data processing. It contains a very extensive set of commands for performing standard mathematical (numerical) operations on matrices and allows the construction of new commands based on the existing ones using an expression parser that interprets command sequences. It also has sophisticated graphics capabilities and provisions for the exchange of data with any user program. This last feature enables MATLAB to communicate with the data processing system over Ethernet by way of a "socket" type communications procedure that is standard in the UNIX operating system. MATLAB thus greatly facilitates development of the monitoring system.

The data processing system is based on a real-time operating system called VxWorks [2] which provides multitasking support and network services. It uses the same file server as the SUN workstation and allows direct downloading of code compiled on the SUN workstation into the real-time system, linking downloaded routines through the use of symbol tables. This arrangement results in a highly efficient environment for developing and testing the data processing software.

The software hierarchy is shown schematically in figure (4.10). Figure (4.11) shows an example of a simple data processing program with a flow diagram on top and the actual code below. *Squares indicate stored data and circles* represent operations. Two processors can work on this problem in parallel.

Simultaneous processing of parallel legs in a single program is implemented by a simple bean passing scheme that prevents data from being overwritten before all successors have used it and prevents execution of operations until all prerequisite data is available.

4.10 APPLICATION TO PRECISION MACHINE TOOL

The methodology presented here will be demonstrated with an example from the Stanford Diamond Turning Machine. The systematic approach to the monitoring system construction which the presented method offers is shown before the example, step by step.

1. *Select parameters.* Identify those parameters that are both critical to the performance of the machine and unknown or subject to change during the machine's operation. When many equivalent parametrizations are possible, the one should be selected that is most meaningful to human

[1] MATLAB is a matrix calculation program produced by The MathWorks Inc. It offers easy data manipulation and versatility along with excellent graphics capabilities.

[2] VxWorks is produced by Wind River Systems Inc.

designers or operators, also considering the use of the parameters in a model. Parameters that are unknown at the outset but not likely to change can be included once and estimated once and for all or they can be estimated off line.

2. *Derive models.* The model should relate the parameters in question to the measurements that are available. Both analytical and empirical models are useful and may be arrived at in any conceivable way. As a part of this research, a detailed model was derived that relates pressures in the pockets and at the measurement locations to the axial forces acting on the rotor. The model accounts for the trapped air and a number of other abnormalities.

3. *Select pattern.* This is where experience and good understanding of the machine is most helpful. The criteria for selecting a pattern are the following:

   - The pattern should be sensitive to one or more of the parameters in question.
   - A model should be able to account for the most prominent features of the pattern.
   - It should be possible to evaluate the variance in the pattern by means of analysis or compilation of statistical data on a set of equivalent patterns. If the model does not account for all the features of the pattern, the resulting biases must be allowed for by increasing the pattern's variance correspondingly.
   - In general, all patterns that depend on the parameter in question should be used in estimating its value. This however may become unfeasible because each pattern may bring into the estimation a number of other unrelated parameters, making the estimation computationally bothersome.
   - Considering the above, it is highly desirable to devise the patterns so that they depend on as few parameters as possible. The best possible pattern is thus one that is sensitive to one parameter only and is therefore simply a meter or a gage of that parameter.

   It is important to note that there are no restrictions on the source or nature of the information behind the pattern, so long as the above criteria are met.

4. *Select data processing procedure.* There may be many ways to generate the pattern from the data. Any method from signal processing or numerical analysis may be used, with due consideration of bias errors and variance.

For this example, we wish to monitor air that tends to get trapped at a certain location in a hydrostatic thrust bearing that constrains axial motion of the turning machine spindle. Analysis has shown that excessive amounts of trapped air render the bearing as much as ten times more compliant in a certain frequency range than it is statically, compromising the performance of the machine.

4.10.1 Parameter Selection

In a detailed model of the thrust bearing that has been derived as part of this research, the amount of air trapped at the sensor port on each side is directly quantified by its volume, V_{a1} at the front end and V_{a2} in the back. Experience shows that success at letting out the air is very erratic so that prior knowledge of how much air there is on each side is very little. It seems impossible to get all the air out (some is always left in such places as fitting threads) whereas the maximum is about $5 \times 10^{-7}\,\mathrm{m}^3$. Most often, the volume of trapped air after a good attempt to let it out is found to be about $2 \times 10^{-8}\,\mathrm{m}^3$ and always more than $1 \times 10^{-9}\,\mathrm{m}^3$. We can assume that it has been attempted to let the air out so that the larger values are rather unlikely.

4.10.2 Pattern Selection

The selection of patterns for detection of trapped air is partly driven by the measurements that are available. In case of the thrust bearing, only the pressure at the pressure measurement port is available. However, measurements are also available of the axial force acting on the bearing during the cutting process. As shown in figure (4.12), the transfer function from external axial force to the pressure as measured at the sensor ports is highly sensitive to the presence of air, especially at the higher frequencies. To get a pattern that best meets the criteria stated in the previous section, it is advisable to stay away from effects that depend strongly on other parameters and of course we must consider only that part of the transfer function model which has been shown to describe the observations accurately, in this case up to about 1 kHz. A segment of the transfer function that meets these requirements is the frequency range from 500 - 900 Hz, i.e. above the effects of the stator dynamics and below the compression mode of the fluid. In order to gain some averaging effect and to minimize the probability of false minima, a set of frequencies is selected (shown with stars on fig. (4.12)) and the transfer function estimated at these frequencies.

4.10.3 Data Processing

Transfer function estimation using discrete Fourier transforms is subject to bias errors when the input signal is corrupted by noise, the system is nonlinear or if the resolution is insufficient to resolve narrow peaks or notches. Since none of these are known to affect the measurement and the frequency range selected for the pattern, the pattern is assumed to be unbiased. Variance in the pattern is simply estimated from the collection of transfer function estimates that are used to form the averaged pattern.

4.10.4 Estimation Results

The following MATLAB dialog shows how the search converges. The two columns of numbers show the t-statistic for each parameter and below is the final parameter value. For the sake of demonstration, we choose to specify prior information as the logarithm (base 10) of the air volume, since the prior knowledge described earlier about air left after bleeding is more about the order of magnitude rather than the number itself. Hence we say that the expected value is -7.7 ($= \log_{10}(2 \times 10^{-8})$) with standard deviation of .5, meaning that we expect values between a third and three times the average to occur most (63%) of the time. The calculation results shown below are based on a pattern of three frequency points (500, 700 and 900 Hz) in the transfer function estimate from axial force to pressure response in the front pocket and same three for the back pocket, resulting in a pattern vector of six elements. The t-vector then contains the normalized statistics for the logarithm of air volume.

```
>> mon
t=
  -2.8009e-01   8.3478e-02
  -3.9602e-01   2.0206e-02
  -6.9303e-01  -2.7720e-01
  -6.3242e-01  -2.8901e-01
  -6.4038e-01  -2.8735e-01
  -6.3957e-01  -2.8752e-01
  -6.3966e-01  -2.8750e-01
  -6.3965e-01  -2.8750e-01
Name    Value       Std. Dev. Unit
---------------------------------
logVa1 -8.02        0.01251   []
logVa2 -7.844       0.01416   []
>>
```

Standard deviations shown are found by calculating eigenvalues of the Fisher information matrix. Observing the pattern has reduced the prior standard deviation thirtyfold and we conclude that a good estimate of the air volumes sought has been found. Figure (4.13) shows the modeled transfer functions that correspond to the most likely fit to the pattern, shown with '+' and '*' for front and back bearing pockets respectively.

4.11 CONCLUSIONS

The research reported here has resulted in a new, systematic approach to machine monitoring that is generally applicable. It allows retrieval of essential monitoring information in a meaningful form with indication of the reliability of the information. It also allows the combination of all relevant data or other quantifiable information in a systematic way to arrive at reliable conclusions. An efficient algorithm has also been derived that carries out parameter estimation under the above scheme. Furthermore, strategies for decisions based on the retrieved information have been investigated.

All the above results are based on extensive prerequsite work that includes detailed modeling of imortant elements of the Stanford Diamond Turning Machine, procurement and installation of computer hardware and instrumentation and development of suitable software. The methodology presented was developed after extensive evaluation of conventional monitoring methods.

The methodology described has been successfully applied to the Stanford Diamond Turning Machine with an example shown in the report.

Bibliography

[1] A. M. Agogino and S. Russell. Sensor fusion using influence diagrams and reasoning by analogy: Applications to milling machine monitoring and control. In J. S. Gero, editor, *Artificial Intelligence in Engineering: Diagnosis and Learning*, pages 333–357. Computational Mechanics Publications, 1988.

[2] Julius S. Bendat and Allan G. Piersol. *Engineering applications of correlation and spectral analysis*. John Wiley and Sons, 1980.

[3] M. P. Boyce, R. K. Bhargava, and R. Chinoy. On-line condition monitoring of turbomachinery. In *Proceedings of the First International Machinery Monitoring and Diagnostics Conference, Las Vegas, NV*, September 1989.

[4] S. Braun, E. Lenz, and C. L. Wu. Signature analysis applied to drilling. *ASME Journal of Mechanical Design*, 104, April 1982.

[5] S. Braun and B. Seth. Analysis of repetitive mechanism signatures. *Journal of Sound and Vibration*, 70(4):513–526, 1980.

[6] S. Braun and B. B. Seth. On the extraction and filtering of signals acquired from rotating machines. *Journal of Sound and Vibration*, 65(1):37–50, 1979.

[7] Simon Braun. The signature analysis of sonic bearing vibrations. *IEEE Transactions on Sonics and Ultrasonics*, SU-27, November 1980.

[8] Arthur E. Bryson, Jr. and Yu-Chi Ho. *Applied optimal control*. Hemisphere Publishing Corp., 1975.

[9] C. Cempel. Diagnostiacally oriented measures of vibracoustical processes. *Journal of Sound and Vibration*, 73(4):547–561, 1980.

[10] E. N. Diei and D. A. Dornfeld. A model of tool fracture generated acoustic emission during machining. *ASME Journal of Engineering for Industry*, 109, August 1987.

[11] N. E. Diei and D. A. Dornfeld. Acoustic emission from the face milling process - the effects of process variables. *ASME Journal of Engineering for Industry*, 109, May 1987.

[12] Hugh F. Durrant-Whyte. Sensor models and multisensor integration. *The International Journal of Robotics Research*, 7(6), December 1988.

[13] K. Eman and S. M. Wu. A feasibility study of on-line identification of chatter in turning operations. *ASME Journal of Engineering for Industry*, 102, November 1980.

[14] C. Evans. Precision engineering: an evolutionary perspective. Master's thesis, Cranfield Institute of Technology, College of Manufacturing, March 1987.

[15] Janos J. Gertler. Failure detection and isolation in complex process plants. A survey. *Proceedings of the IFAC Symposium om Microcomputer Applications to Process Control, Istanbul, Turkey*, pages 13–25, 1986.

[16] Steven Ray Hall. *A Failure Detection Algorithm for Linear Dynamic Systems*. PhD thesis, Massachusetts Institute of Technology, June 1985.

[17] H. D. Haynes and R. C. Kryter. Condition monitoring of machinery using motor current signature analysis. In *Proceedings of the First International Machinery Monitoring and Diagnostics Conference, Las Vegas, NV*, September 1989.

[18] Tom Henderson and Esther Shilcrat. Logical sensor systems. *Journal of Robotic Systems*, 1(2), 1984.

[19] Yu-Chi Ho and Ashok K. Agrawala. On pattern classification algorithms. *IEEE Transactions on Automatic Control*, AC-13, December 1968.

[20] M. S. Hundal. Mechanical signature analysis. *Shock and Vibration Digest*, 15, June 1983.

[21] Rolf Isermann. Process fault detection based on modeling and estimation methods — a survey. *Automatica*, July 1984.

[22] Elijah Kannatey-Asibu and D. A. Dornfeld. Quantitative relationships for acoustic emission from orthogonal metal cutting. *ASME Journal of Engineering for Industry*, 103, August 1981.

[23] S. M. Kay and S. L. Marple. Spectrum analysis: A modern perspective. *Proceedings of the IEEE*, 69, November 1981.

[24] Steven M. Kay. *Modern spectral estimation*. Prentice Hall, 1987.

[25] Thomas H. Kerr. A critique of several failure detection approaches for navigation systems. *IEEE Transaction on Automatic Control*, 34(7), july 1989.

[26] G. B. Kliman et al. Broken bar detector for squirrel cage induction motors. In *Proceedings of the First International Machinery Monitoring and Diagnostics Conference, Las Vegas, NV*, September 1989.

[27] Ren C. Luo and Michael G. Kay. Multisensor integration and fusion in intelligent systems. *IEEE Transactions on Systems, Man and Cybernetics*, 19(5), September/October 1989.

[28] R. H. Lyon and A. Ordubadi. Use of cepstra in acoustical signal analysis. *ASME Journal of Mechanical Design*, 104, April 1982.

[29] Richard H. Lyon. *Machine Noise and Diagnostics*. Butterworths, 1987.

[30] Richard H. Lyon. The monitoring and diagnostics of machines and processes. In *Proceedings of the First International Machinery Monitoring and Diagnostics Conference, Las Vegas, NV*, September 1989.

[31] R. K. Mehra and J. Peschon. An innovations approach to fault detection and diagnosis in dynamic systems. *Automatica*, 7:637–640, 1971.

[32] Arye Nehorai. A minimal parameter adaptive notch filter with constrained poles and zeroes. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-33, August 1985.

[33] A. V. Oppenheim and R. W. Schafer. *Digital Signal Processing*. Prentice-Hall, Inc., 1975.

[34] S. M. Pandit and S. Kashou. A data dependent systems strategy of on-line tool wear sensing. *ASME J. of Engineering for Industry*, August 1982.

[35] S. M. Pandit, T. L. Subramanian, and S. M. Wu. Modeling machine tool chatter by time series. *ASME Journal of Engineering for Industry*, February 1975.

[36] S. M. Pandit, H. Suzuki, and C. H. Kahng. Application of data dependent systems to diagnostic vibration analysis. *ASME Journal of Mechanical Design*, 102, April 1980.

[37] Lois-Francois Pau. *Failure diagnosis and monitoring*. Marcel Dekker, Inc., 1981.

[38] V. F. Pisarenko. The retrieval of harmonics form a covariance function. *Geophys. J. R. astr. Soc*, 33:347–366, 1973.

[39] A. Rault, D. Jaume, and M. Vergé. Industrial process fault detection and localization. In *Proceedings of the 9th IFAC World Congress, Budapest, Hungary*, july 1984.

[40] Ross D. Shachter. Evaluating influence diagrams. *Operations Research*, 34(6), November/December 1986.

[41] Petre Stoica and Arye Nehorai. The poles of symmetric lineאr prediction models lie on the unit circle. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-34, October 1986.

[42] Tad J. Ulrych and Rob W. Clayton. Time series modelling and maximum entropy. *Physics of the Earth and Planetary Interiors*, 12:188–200, 1976.

[43] Richard J. Vaccaro. On adaptive implementation of pisarenko's harmonic retrieval method. *Proceedings of the ICASSP, San Diego, Calif., March 19.-21.*, 1984.

[44] H. L. van Trees. *Detection, Estimation and Modulation Theory*, volume 1. McGraw-Hill, 1968.

[45] R. H. Volin. Techniques and aplications of mechanical signature analsysis. *Shock and Vibration Digest*, 11, September 1979.

[46] J. Wagner and R. Shoureshi. A robust failure diagnostics scheme for nonlinear thermofluid processes. In *Proceedings of the American Control Conference, Minneapolis*, 1987.

[47] B. K. Walker. Recent developments in fault diagnosis and accomodation. *Proceedings of AIAA Guidance and Control Conference, Gatlinburg, Tenn., August*, 1983.

[48] Alan S. Willsky. A survey of design methods for failure detection in dynamic systems. *Automatica*, 12:601–611, 1976.

[49] Alan S. Willsky and Harold L. Jones. A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems. *IEEE Transactions on Automatic Control*, February 1976.

[50] S. M. Wu, T. H. Jr. Tobin, and M. C. Chow. Signature analysis for mechanical systems via dynamic data systems (dds) monitoring technique. *ASME J. of Mechanical Design*, 102, April 1980.

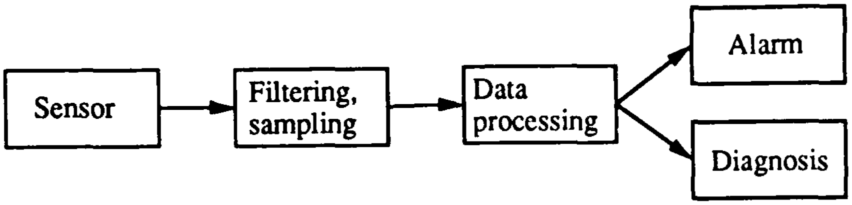


Figure 4.1: Typical monitoring system setup. A single signal is sampled and processed in a single computer.

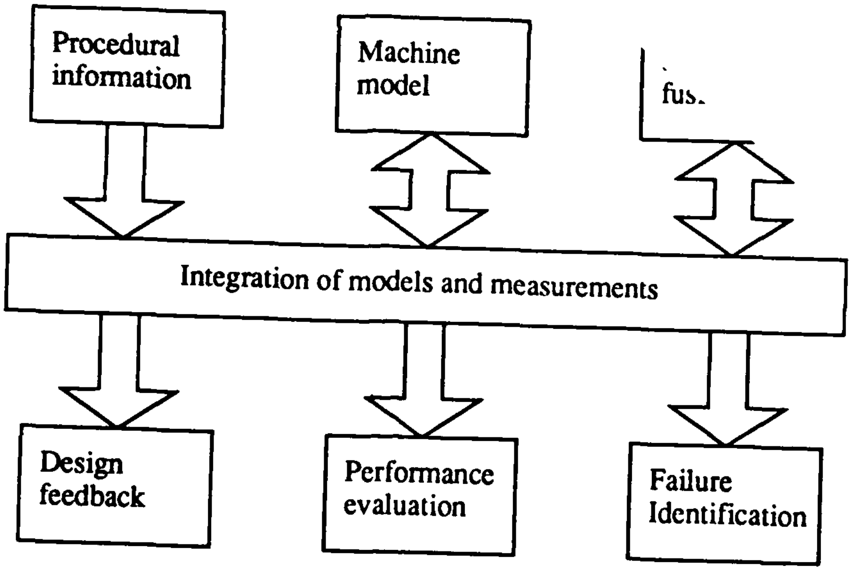


Figure 4.2: This figure shows schematically the most important parts of a monitoring system and how they interact.

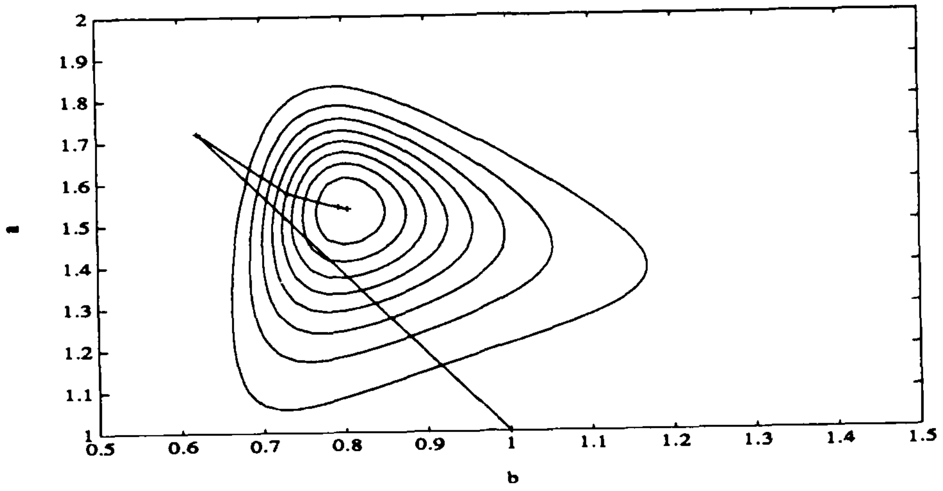


Figure 4.3: Level curves of posterior density with convergence path shown. Prior densities as specified in text. Model is: $m(\theta) = [\theta_1^3\theta_2, \theta_1/\theta_2^3]^T$.

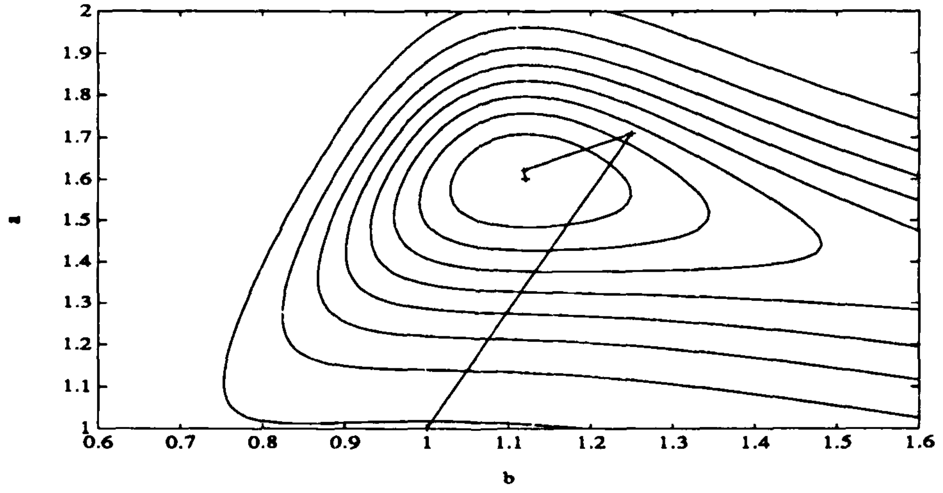


Figure 4.4: Level curves of posterior density with convergence path shown. Prior densities as specified in text. Model is: $m(\theta) = [\theta_1^2\theta_2, \theta_1^2/\theta_2^3 + \theta_2]^T$.

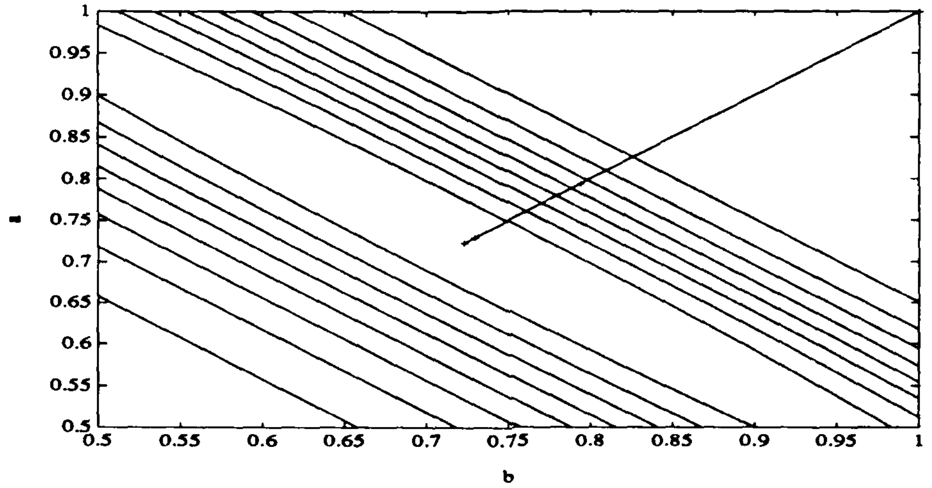


Figure 4.5: Degenerate model, i. e. it has only one degree of freedom which is a linear combination of the parameters: $m(\theta) = [(\theta_1 + \theta_2)^3, (\theta_1 + \theta_2)^{-3}]^T$.

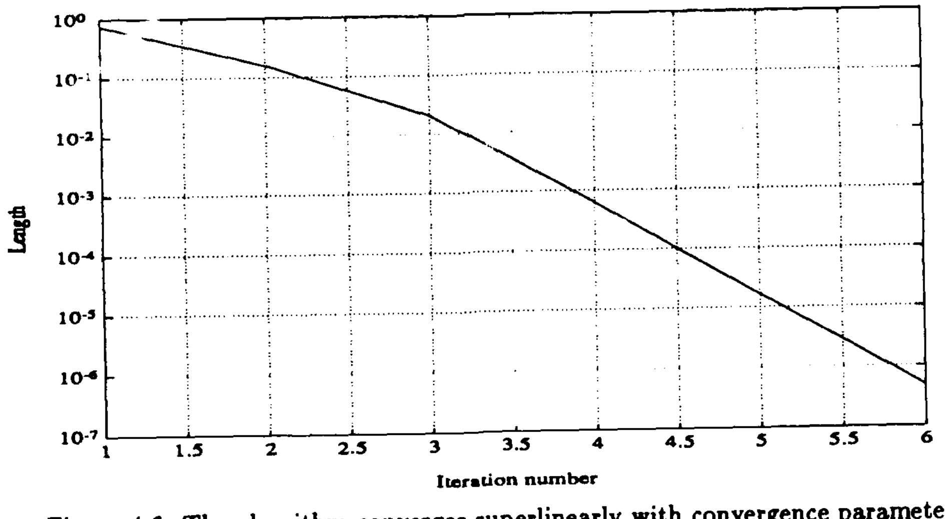


Figure 4.6: The algorithm converges superlinearly with convergence parameter in this case of .03 per iteration.

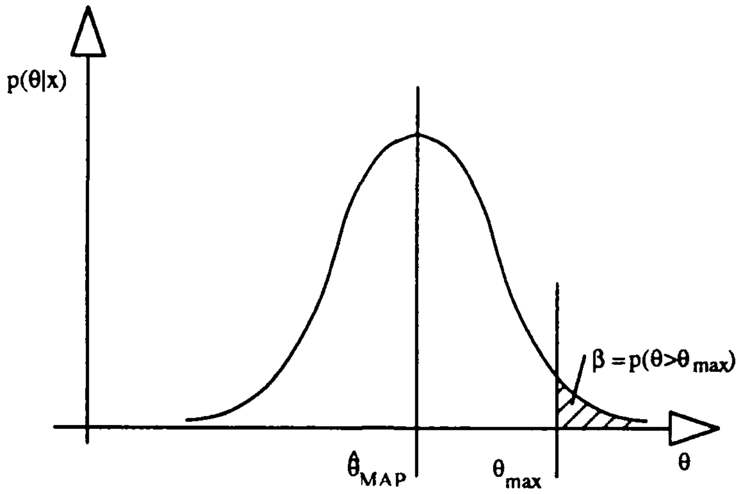


Figure 4.7: If the density of the posterior estimate is known, the probability of misclassification (β) can be calculated.

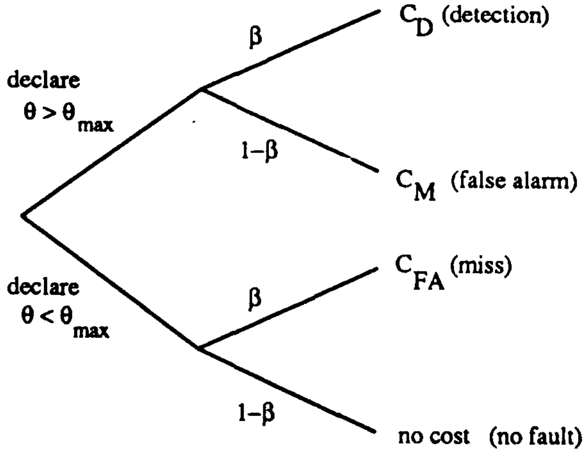


Figure 4.8: Decision tree for the classification problem. Each decision (declaring or not declaring θ over the limit) can be either right or wrong. The best strategy is to issue the declaration that is expected to have the lowest cost.

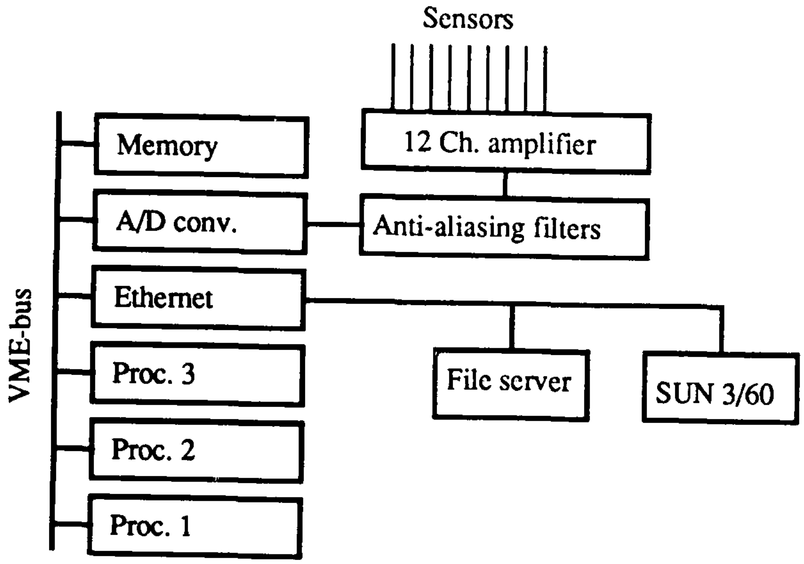


Figure 4.9: This figure shows schematically interconnections between the computer hardware modules and the instrumentation.

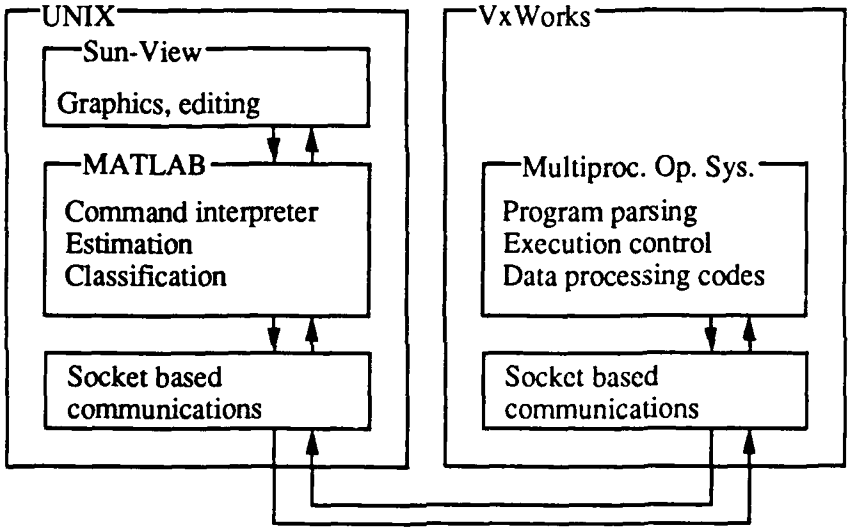
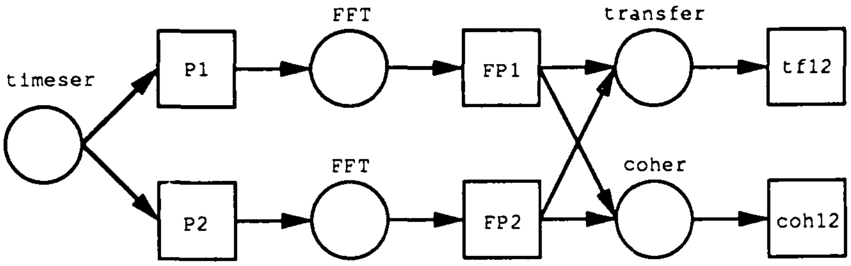


Figure 4.10: This diagram shows software modules at the highest level and how they are nested. Unix is the operating system for the SUN-workstation and VxWorks runs on the real-time system.



```
% matrix declaration syntax:
% <type>       <name>   <len> <width> *<chan>   <sens>   <ampl>   <unit>   <xdcr>

double matrix P1       1024  1        *0        6.7e-6    1      Pascal   7563
double matrix P2       1024  1        *1        6.7e-6    1      Pascal   7563
double matrix FP1      1024  1
double matrix FP2      1024  1
double matrix tf12     1024  1
double matrix coh12    1024  1

net
% description of the net

P1, P2 = timeser( 1024.0, 16000.0 )
FP1 = FFT( P1 )
FP2 = FFT( P2 )
tf12 = transfer( FP1, FP2 )
coh12 = coher( FP1, FP2 )
```

Figure 4.11: The figure shows a flow diagram corresponding to the program listing below. Squares are data modules and circles represent operations.

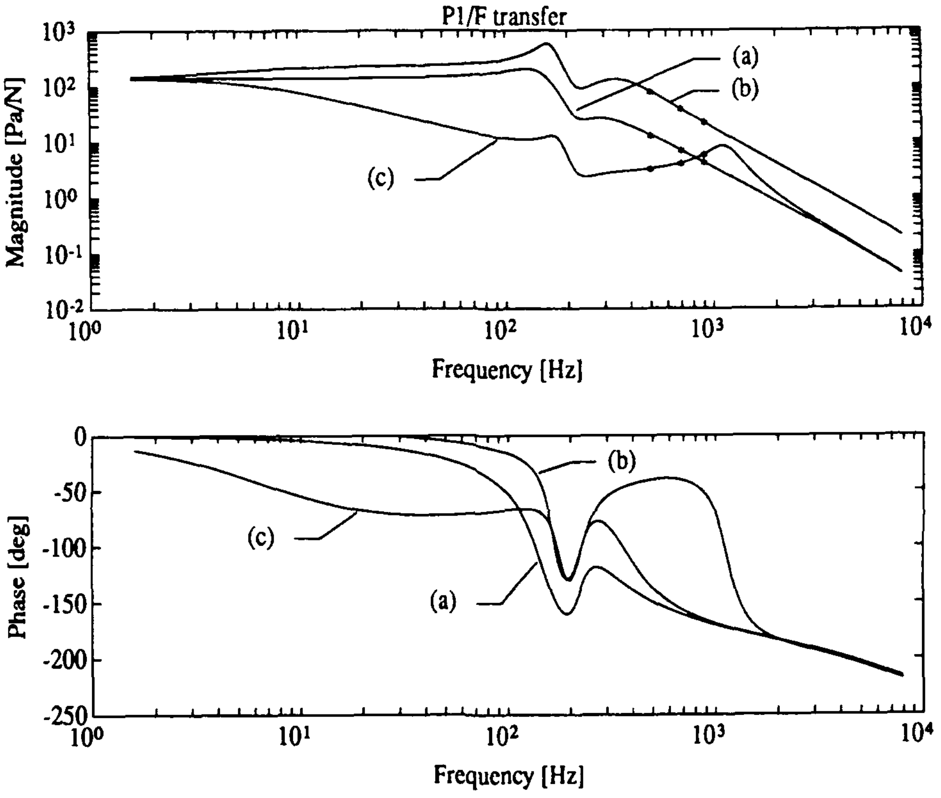


Figure 4.12: The figure shows dependency of the $P_{m1}(s)/F(s)$ transfer function on the amount of air trapped in the front pocket. Graphs shown assume following amounts in front pocket: (a) .5cm^3, (b) .1cm^3 and (c) negligible amount. In all cases, .5cm^3 are assumed in back pocket. Stars indicate frequencies selected for pattern.

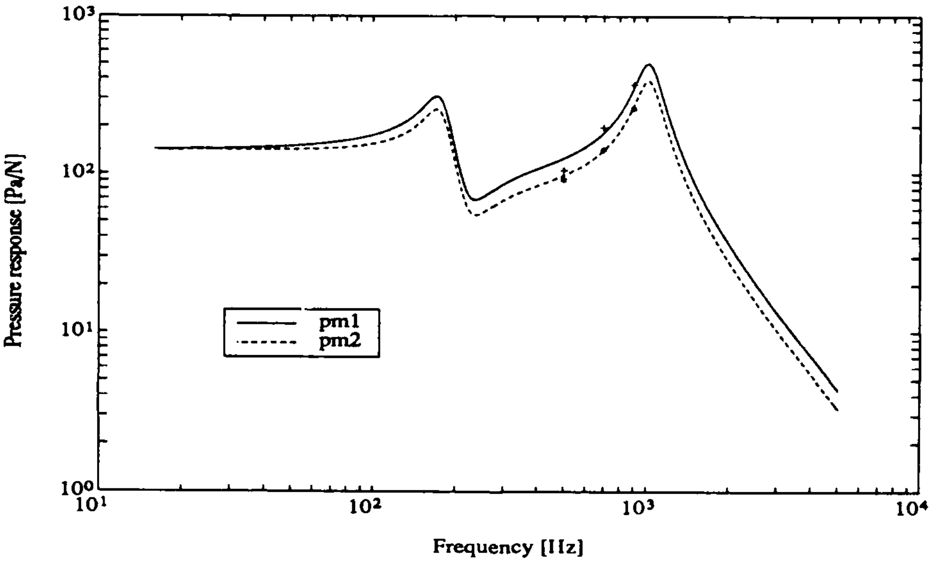


Figure 4.13: The solid and dashed lines show the transfer from axial force to pressure response at measurement port for front and back bearing halves respectively, as projected by the model. Air volume has been set to most probable fit to the pattern which is indicated with '+' marks for front and '*' for back.